frontiers in Genetics

Leveraging user-friendly network approaches to extract knowledge from high-throughput omics datasets

Pablo I. Ramos[1*], Luis W. Arge[2], Nicholas C. Lima[3], Kiyoshi F. Fukutani[4], Artur T. de Queiroz[1]

[1]Center for Data and Knowledge Integration for Health (CIDACS), Gonçalo Moniz Institute (IGM), Brazil, [2]Laboratório de Genética Molecular e Biotecnologia Vegetal, Centro de Ciências da Saúde, Federal University of Rio de Janeiro, Brazil, [3]Departamento de Bioquímica e Biologia Molecular, Universidade Federal do Ceará, Brazil, [4]Multinational Organization Network Sponsoring Translational and Epidemiological Research (MONSTER) Initiative, Fundação José Silveira, Brazil

Submitted to Journal:
Frontiers in Genetics

Specialty Section:
Systems Biology

ISSN:
1664-8021

Article type:
Review Article

Received on:
29 Mar 2019

Accepted on:
16 Oct 2019

Provisional PDF published on:
16 Oct 2019

Frontiers website link:
www.frontiersin.org

Citation:
Ramos PI, Arge LW, Lima NC, Fukutani KF and De_queiroz AT(2019) Leveraging user-friendly network approaches to extract knowledge from high-throughput omics datasets. *Front. Genet.* 10:1120. doi:10.3389/fgene.2019.01120

Leveraging user-friendly network approaches to extract knowledge from high-throughput *omics* datasets

Pablo Ivan Pereira Ramos[1*], Luis Willian Pacheco Arge[2], Nicholas Costa Barroso Lima[3], Kiyoshi F. Fukutani[4], Artur Trancoso L. de Queiroz[1]

Affiliations:

[1] Center for Data and Knowledge Integration for Health (CIDACS), Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador, Brazil
[2] Laboratório de Genética Molecular e Biotecnologia Vegetal, Centro de Ciências da Saúde, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
[3] Departamento de Bioquímica e Biologia Molecular, Universidade Federal do Ceará, Fortaleza, Brazil
[4] Multinational Organization Network Sponsoring Translational and Epidemiological Research (MONSTER) Initiative, Fundação José Silveira, Salvador, Brazil.

***Correspondence:**

Dr. Pablo Ivan Pereira Ramos

pablo.ramos@fiocruz.br

Number of words in the Abstract: 154

Number of figures: 6

Number of tables: 5

Author's e-mail addresses:

AQ: artur.queiroz@fiocruz.br

KF: ferreirafk@gmail.com

LA: l.willianpacheco@yahoo.com.br

NL: ncblima@gmail.com

PR: pablo.ramos@fiocruz.br

Abstract

Recent technological advances for the acquisition of multi-*omics* data have allowed an unprecedented understanding of the complex intricacies of biological systems. In parallel, a myriad of computational analysis techniques and bioinformatics tools have been developed, with many efforts directed towards the creation and interpretation of networks from this data. In this review, we begin by examining key network concepts and terminology. Then, computational tools that allow for their construction and analysis from high-throughput *omics* datasets are presented. We focus on the study of functional relationships such as co-expression, protein-protein interactions, and regulatory interactions that are particularly amenable to modeling using the framework of networks. We envisage that many potential users of these analytical strategies may not be completely literate in programming languages and code adaptation, and for this reason, emphasis is given to tools' user-friendliness, including plugins for the widely adopted Cytoscape software, an open-source, cross-platform tool for network analysis, visualization, and data integration.

Keywords

correlation networks; graph; high-throughput sequencing; network analysis; omics; protein-protein interaction; regulatory networks; systems biology.

1. Introduction

The analysis of high-throughput datasets using the framework of networks has gained widespread adoption in the biological sciences. With approaches in this field shifting from a mostly reductionist perspective towards a more holistic view of natural phenomena (Barabási and Oltvai, 2004; Berlin et al., 2017), the analytical tools used to extract knowledge from data have also adapted. The vocabulary of networks is particularly suitable for studying problems that explicitly focus on the *relationships* among elements, where the latter can be any entity under study, including but not limited to genes, transcripts, proteins, or metabolites. With sheer amounts of data that can be obtained from instruments such as high-throughput sequencers, analytical strategies that permit broader insights of the functional roles of each element are warranted, and this can be achieved by the use of network approaches.

In this Review, we focus on the various uses of network methods to the analysis of large-scale *omics* datasets, which are those generated using medium- and high-throughput

technologies in genomics, transcriptomics, proteomics, and metabolomics experiments. First, key concepts and terminology of this area are presented, followed by the introduction of biological network variants, namely correlation networks (Section 2.1), gene regulatory networks (Section 2.2), and protein-protein interaction networks (Section 2.3). Methods to perform key analysis in a network are presented in Section 2.4. With every approach, computational tools that we considered both appropriate and user-friendly are presented. User-friendly tools were defined as those that provide a point-and-click graphical user interface (GUI), which does not mean that they have limited functionality or that they are only used by those without extensive programming literacy. Rather, they can be used to complement analyses performed in different environments, such as R or Python scripts, and usually offer improved layouts and visualization modes compared to less friendly alternatives. Our Review differs from that of others who have engaged in similar challenges (for instance, the works of Aittokallio and Schwikowski, 2006; Huang et al., 2017; and Stevens et al., 2014), since we primarily target the non-programmer who wants to apply network methods to a dataset of interest. Luckily, network analysis is an area that has greatly benefited from the existence of excellent analysis software such as Cytoscape (Shannon et al., 2003) (https://cytoscape.org/), Gephi (Bastian et al., 2009) (https://gephi.org), and NAViGaTOR (Brown et al., 2009), to name a few. Gephi and Cytoscape, in particular, can be extended by the many plugins created by third-party developers and available in official repositories (Saito et al., 2012), and these were at the heart of the current review. While the aforementioned types of networks are widely employed, there are many other applications that are not in the scope of this work. As an example, the modeling of (bio)chemical networks using graph-theoretic approaches have advanced our understanding of bacterial and eukaryotic metabolism (Dutta et al., 2014; Jha et al., 2015; Klein et al., 2012), and were the object of previous reviews (see, e.g., Cottret and Jourdan, 2010; Lacroix et al., 2008). Biology and Biomedicine are, indeed, areas which have been greatly benefited by the use of network techniques resulting from cross-pollination among disciplines.

1.1. Beyond the empirical, towards formalism: what are networks?

Network is a general term used in many different contexts: social networks, traffic networks, ecological networks, computer networks, among others, all share a common theme related to the interaction among a set of disparate elements, viz. people, vehicles, species, and computers. The topology of networks and the interactions within can be formally studied from a graph-theoretic viewpoint, which allows for a mathematical representation and

formalism, while also facilitating visualization of the network. Since several distinct graph representations exist, for generality we will focus on the description of simpler types of graphs. In general, a graph $\mathcal{G} = (V, E)$ is composed of a finite set V of nodes (or vertices), and E of (directed or undirected) edges (or links). In the case of *omics* datasets, each node $v \in V$ could represent a (bio)chemical entity such as a gene, transcript, protein, or metabolite, and an edge $e = \{v_1, v_2\} \in E$ exists between two nodes when there is evidence for their interaction, which in turn depends on the specific aim of the modeled network, which guides the definition of interaction. For instance, in the simplest type of correlation network, one could specify a hard threshold over all pairwise values of Pearson's correlation coefficients in order to determine whether any two nodes are connected. On the other hand, in a protein-protein interaction (PPI) network, edges between protein nodes exist when evidence for their physical interaction is available, which could be obtained by a wealth of techniques that include co-immunoprecipitation, affinity purification, proteomics, and computational approaches (Ngounou Wetie et al., 2014).

The edges in a graph can be undirected (**Figure 1A**) or directed (**Figure 1B,C**), depending on whether the interactions between elements are symmetric or not. In directed graphs, there is a specific sense pointing at the direction of a given interaction, such as a transcription factor that regulates a given gene in a regulatory network (a causal relationship), while undirected graphs describe two-way associations such as the co-expression of genes in a correlation network, in which a significant correlation *per se* does not provide sufficient evidence to infer whether any of the compared genes regulates or is being regulated by the other, or even by an upstream regulator acting on both simultaneously. That is, correlation does not imply causation, and hence the undirected graph is a more appropriate representation of this relationship.

Graphs can also have numerical weights associated with each interaction, the interpretation of which depends on the specific application under study (**Figure 1C**). In a correlation network, for instance, weights could represent the magnitude of the correlation statistic. Also possible is to set weights based on the confidence of the interaction as measured by a relevant parameter. As an example, the STRING database (http://string-db.org), which harbors information on physical and functional protein-protein interactions, quantifies interaction weights between proteins as a combined score dependent on the nature (experimental or

computational prediction) and quality of the supporting evidence (Szklarczyk et al., 2017). **Table 1** summarizes the biological interpretation of nodes, edges, and edge weights for the three types of networks considered in this study. While these interpretations are typical for these kinds of biological networks, studies may employ different analytical strategies that lead to variations on how to account edge directionality or weights, for instance. As an example, regulatory networks are usually inferred using a bipartite graph representation, where nodes are of two different types (either a transcription factor or a target gene). In this case, edge directionality characterizes an underlying regulatory event (activation or inhibition) of a transcription factor towards a target gene, hence these networks are usually modeled as a directed graph (Narasimhan et al., 2009; Song et al., 2017).

2. How to disclose networks from high-throughput *omics* datasets

In the following sections, we review and discuss methods to construct various types of networks using a wealth of *omics* datasets as input (**Figure 1D**). While many different computational methodologies to achieve the construction of a network exist, we focus on those that we considered more apt for users without a computational background, especially those that are based on plugins for the popular software Cytoscape (Shannon et al., 2003), which allows visualization, rendering, and analysis of networks in the same computational environment, with the advantage of being open-source, platform-independent, and continuously updated. Once the tools to build these biological networks are covered, we shift our focus towards analysis and visualization aspects of graphs, which are covered in Section 2.4 (**Figure 1E**).

2.1. Correlation networks allow disclosing of relevant associations in *omics* datasets

Recent advances in high-throughput technologies have increased our capacity to assess the elements in different *omics* layers, allowing their simultaneous treatment in single grouped mechanisms that together explain biological events (Carpenter and Sabatini, 2004; Vella et al., 2017). In this sense, the processes that allow for life maintenance in cells can be regarded as an intricate web of complex relationships between molecules such as proteins, lipids, metabolites, and nucleic acids (RNA and DNA) (Barabási et al., 2011). Correlations are arguably the dominant way to infer relationships not only between the elements in these distinct layers of information but also within each layer, as it allows simultaneously examining the associations that drive an observed biological effect, and there are several ways of calculating correlation coefficients. Statistically, the correlation is a measure of the

two-way linear association between a pair of variables (Mukaka, 2012). The correlation coefficient permits estimating the degree or strength of this association. The most common and classic correlation statistic is the Pearson's correlation coefficient (or *r*), which measures linear associations between two variables under the assumption that the data be normally distributed and that observations are independent (Walter and Altman, 1992). Non-parametric methods based on ranks avoid the assumption of normality and are preferred when the data is ordinal, skewed, or presents extreme values (outliers). One such method is the Spearman correlation coefficient, which is a calculation of Pearson's correlation coefficient on the ranks of the observations, rather than on the raw data, and yields an r_S statistic (also called ρ, rho). The Kendall rank correlation coefficient (also called τ, tau) uses the number of concordant and discordant rank pairs to evaluate association. The biweight midcorrelation is less prone to outlier influence because it is a median-based estimation and, like the two previous, yields a robust measurement of association, with the drawback that few tools are available that calculate this metric (Langfelder and Horvath, 2012). Correlation coefficients (r, r_s, ρ, or τ) are a dimensionless quantity ranging from -1 to 1, where values close to zero indicate no (linear) association whilst values equal to or near 1 (or -1) indicate strong, positive (or negative) correlations, although absolute values as low as 0.3 can already be considered a weak correlation depending on the context (Mukaka, 2012).

Since the relationships between genes, proteins, metabolites and biological entities in general are complex and often nonlinear, while having distributions that can be non-normal, alternative measurements of association are often required (Hardin et al., 2007), and include information-theoretical measures such as Mutual Information (MI). MI quantifies the dependence between a pair of random variables and, based on the concept of entropy, estimates how much knowledge is gained about a variable (say, expression values of a gene *X*) by observing a second variable (say, expression values of a gene *Y*), hence its name. The MI is zero when the variables are statistically independent, while a positive value denotes a degree of dependence (Steuer et al., 2002). In a scenario of statistical independence, the distribution of values of variable *X* is not altered at all when those of variable *Y* changes. It is worth noting that traditional association measures that disclose only linear relationships are insufficient to reveal statistical independence, exactly because there can be non-linear relationships in the data that these methods do not adequately capture. We refer the reader to the review of de Siqueira Santos et al. (2014) on statistical dependency identification, who

further provide illustrative biological examples and simulations using various association statistics.

Correlations can be visually assessed by plotting the data as a scatter plot fitted by a line, where the further the data lie from the straight line, the weaker the correlation (**Figure 2A**). While this approach is feasible when few variables are compared, it has limited practicality when dealing with large-scale *omics* datasets, such as high-throughput expression profiling and proteomics. In these cases, methods that create correlation networks are preferred (Langfelder and Horvath, 2008; Vella et al., 2017; Zhang and Horvath, 2005). Once a correlation (or other association statistic) matrix is attained, a network can be inferred. A co-expression network is a particular case of correlation network constructed using genome-wide expression data, although the term is sometimes used to refer to networks created by correlating the abundance of protein or metabolites in proteomics and metabolomics studies. In this network, the nodes are elements such as genes, proteins, or metabolites, and an undirected edge connects a pair of nodes if the correlation statistic between them exceeds a given threshold (**Figure 2C**). This 'hard-threshold' approach represents the simplest form of inducing a network from *omics* data, and is limited by the arbitrary nature of the threshold used, which will dismiss slightly undervalued correlations that could be potentially relevant. An alternative, more sophisticated approach to disclose co-expression networks is by using soft-thresholding approaches, of which the weighted gene co-expression network analysis (WGCNA) algorithm is among the most widely employed methods (Langfelder and Horvath, 2008). The main advantage of the WGCNA approach is that no arbitrary thresholding on the correlation values is enforced, which effectively preserves the continuous nature of the correlation distribution. In addition, it is not impacted by the arbitrariness of hard-thresholding methods. In WGCNA, once all pairwise correlations are calculated, an adjacency matrix, which holds information on edge strengths, is obtained by applying a power transformation of the form $f(x) = x^{\beta}$, where x are correlation values and β is the soft-thresholding parameter, a positive value set by the user such that the resulting network presents an approximately scale-free property while maintaining high connectivity (see **Box 1** for a primer of important network definitions). As a result, high correlations are emphasized at the expense of low correlations, but without the need of setting an explicit threshold on the correlation values themselves.

User-friendly tools for constructing correlation networks

Gene/protein correlation network analysis can be performed using in-house scripts and packages for general-purpose programming languages such as R, Python, Perl, or Java. However, alternatives exist for the bioinformatics user that wants to apply such methods to their data in the absence of a solid computational background (**Table 2**). One of them is based on the Cytoscape environment, which also allows for installing third-party plugins. A specific app developed for correlation network analysis, the *ExpressionCorrelation* app (available at http://apps.cytoscape.org/apps/expressioncorrelation), presents a Pearson's correlation-based solution. Thus, a table of gene/protein/metabolites measurements is the input and Cytoscape can generate the gene and sample correlation network. This plugin has been applied to the construction of many networks, exemplified by an *Anopheles* gene co-expression network (Shrinet et al., 2014), a correlation network from *Aspergillus* metabolites highlighting those significantly associated to anticancer and antitrypanosomal bioactivity (Tawfike et al., 2019), and co-expression networks from cancer datasets (Wang et al., 2016b; Zhang et al., 2016). Pearson's correlation statistic, however, presents several limitations as pointed out in the previous section. The Cyni toolbox app circumvents this difficulty by allowing calculation of rank-based correlations such as Spearman's and Kendall's, in addition to Pearson's coefficient (Guitart-Pla et al., 2015). **Figure 3** shows a bacterial co-expression network constructed using Cyni.

Another user-friendly solution is *geWorkbench* (Floratos et al., 2010). This tool is an open source Java desktop application that allows correlation using an ARACNe (mutual information-based) implementation (Margolin et al., 2006a), and is particularly suitable for finding regulatory networks from transcriptomic data. In addition, the workbench allows for parameter estimation and is fairly flexible for user customization. Its advantages over the Cytoscape *ExpressionCorrelation* app include the possibility of p-value threshold modification and correction, as well as bootstrap resampling. Thus, the program permits evaluating the statistical significance of the network and keep the more robust associations. However, the user-friendly advantage is not without its costs: the plugin is limited to the calculation of regular correlations (Pearson's and Spearman's) and mutual information. Also, the use of more robust correlation statistics, such as the biweight midcorrelation, still requires proficiency in programming languages/R packages, since so far there are no alternatives that incorporate this measure.

The construction of weighted networks using the soft-thresholding approach employed by WGCNA requires the execution of a multi-step pipeline implemented as an R package (Langfelder and Horvath, 2008), thus requiring programming skills to correctly adapt and parametrize the functions and the dataset itself. To circumvent this need, a webserver adaptation of the WGCNA method was recently published as *webCEMiTool*, allowing an user-friendly approach to disclose a weighted co-expression network, detect modules therein, and produce publication-quality visualizations (https://cemitool.sysbio.tools/) (Cardozo et al., 2019). In this context, modules are considered as groups of genes with similar expression profiles, which tend to have related biological functions or be under the influence of the same transcriptional regulator, but a more ample discussion of modularity is presented in Section 2.4. *webCEMiTool* also has a built-in method to automatically select the optimal value of β (the soft-thresholding parameter), which is described elsewhere (Russo et al., 2018) and, like the original WGCNA algorithm, it could also be used to disclose correlation networks from proteomics or metabolomics datasets. Pathway enrichment analysis can be run directly from the *webCEMiTool* application, as it interfaces with the Enrichr platform (Kuleshov et al., 2016) which comprises over a hundred gene set libraries, thus facilitating the interpretation and extraction of knowledge from the inferred network.

2.2. Gene regulatory networks permit an improved understanding of the cell's transcriptional circuitry

Gene (transcriptional) regulatory networks, or GRNs, are models that aim at the elucidation of genetic information processing, aiding on the understanding of organism development. A GRN is based on the following elements: transcription factors (TFs), target genes, and their regulatory elements in the upstream region. TFs are identified using computational tools based on sequence homology and through motif conservation across transcription factor families. Each TF can act on the transcription of multiple genes. In the upstream region of each target gene, there exist elements/motifs that are recognized by the TF, and the gene is subsequently transcribed. When located upstream of a gene, these motifs are called *cis*-elements. Identification of *cis*-elements can be performed by biological experiments, such as by chromatin immunoprecipitation (ChIP)-seq methodology (Lee et al., 2006), or computationally by alignment of known motifs or by the identification of novel motifs. The latter are called *de novo* approaches and employ mathematical structures such as hidden Markov models (HMM) (Bailey et al., 2009). Typically, after the identification or discovery

of new *cis*-elements, an enrichment analysis is performed using Fisher's exact test for identification of enriched motifs in the set of upstream regions from target genes.

On the other hand, the prediction of TFs-target genes interactions can be performed using a reverse engineering-based strategy. The top-down approach is particularly suitable in this context and uses information from gene expression datasets to detect expression patterns and then induce a GRN (Hache et al., 2009; Hartemink, 2005). The first models used to infer GRNs were based on the Pearson correlation coefficient but failed to capture non-linear pattern dependencies (as previously addressed). Other approaches were subsequently developed and applied to disclose GRNs in a more robust way, and included regression (Huynh-Thu et al., 2010), mutual information (Margolin et al., 2006a), partial correlations (Wille et al., 2004), and variations of these (Luo et al., 2008; Meyer et al., 2008). Despite each method having its peculiarities, GRNs inferred by diverse techniques usually do not present large differences (de Matos Simoes et al., 2013), and bootstrap analysis could be used to infer more robust GRNs. Another difficulty is the existence of regulation patterns that occur in rare conditions and cannot be easily detected, requiring specific wet-lab experiments for this purpose.

The study of gene regulation can take two main paths: i) GRN inference and ii) dynamic modeling, which can be performed either in isolation or in conjunction. We focused on methods that accomplish the first goal, while the latter can be attained using a diverse array of techniques that include Boolean formalism (logical models), Bayesian dynamic networks, and Ordinary Differential Equations (studied elsewhere, *e.g.*, Kaderali and Radde 2008; Naldi et al. 2009; and Chai et al. 2014). The representation of inferred GRNs can be in the form of bipartite graphs which, in contrast to the simple graphs presented in the Introduction and in the construction of co-expression networks, have nodes of two types: TFs or target genes, and edges between them indicate a regulatory interaction (**Table 1, Figure 4A**). This type of representation is usually employed to GRNs originated from co-expression relationships because usually no *a priori* information is available about the type of regulation that the transcription factor exerts on the target genes. Logical models, on the other hand, incorporate prior information on gene activation and repression, and the modeling of these relationships permit the capturing of the global dynamic behavior of the regulatory network in a simple fashion. An example of such a network from the human GRN, available in TRRUST database, is shown in **Figure 4B**.

User-friendly tools for constructing gene regulatory networks

As seen above, construction of GRNs is based on interaction inference between TFs and target genes, and on the identification of *cis*-elements in the upstream region of target genes. Next, we present user-friendly tools to perform both steps. Gene regulatory networks inferred based on gene expression patterns are considered of intermediate value because they require improvement and validation with biological experiments. Traditionally, the inference of GRNs has been performed with tools based on command-line or in the R programming language such as ARACNe (Margolin et al., 2006a), but current alternatives include more user-friendly approaches which are listed in **Table 3**. These include an ARACNe implementation in *geWorkbench*, which was listed previously in the correlation network section, and also available are the Cytoscape plugins CyGenexpi (Modrák and Vohradský, 2018), CyNetworkBMA (Fronczuk et al., 2015), GRNCOP2 (Gallo et al., 2011), and iRegulon (Janky et al., 2014) (**Table 3**).

The ARACNe package is based on mutual information index to establish interactions between a pair of genes, such as a TF and a target gene; moreover, this tool employs bootstrapping to generate a consensus and robust network (Margolin et al., 2006b). CyGenexpi is based on an ordinary differential equation (ODE) model applied on time series data that together with static binding (*e.g.,* ChIP-seq) or information obtained from the literature allows inferring of gene regulatory modules in bacteria (Modrák and Vohradský, 2018). CyNetworkBMA employs a Bayesian Model Averaging algorithm to infer GRNs with a user-friendly interface and executes network processing on top of R code, which accelerates the inference process by allowing parallel processing (Fronczuk et al., 2015). Additionally, CyNetworkBMA can compute some statistics for the network evaluation, including ROC (Receiver Operating Characteristic) and precision-recall curves. The package GRNCOP2 has an algorithm based on machine learning with a model-free combinatorial optimization to infer time-delayed gene regulatory networks from genome-wide time series datasets (Gallo et al., 2011). The GRNs inference from the iRegulon package is based on analysis of *cis*-regulatory sequences from target genes and performs a genome-wide ranking-and-recovery strategy to detect enriched motifs related to TFs and their optimal sets of direct targets (Janky et al., 2014).

Like other types of biological data, GRNs can be stored on public databases which can be queried by other scientists. In this context, databases that permit storing and downloading of

GRNs include TRRUST (Han et al., 2018), RegNetwork (Liu et al., 2015), ORegAnno (Lesurf et al., 2016), and rSNPBase (Guo and Wang, 2018) (**Table 3**). TRRUST (Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining) database contains information obtained by computational mining and curated TFs-target genes interactions, and about TFs *cis*-regulatory elements in human and mouse. RegNetwork contains information of genic regulations by TFs and microRNAs, also in human and mouse. Similarly, NetworkAnalyst is a webserver that offers an integrated environment to establish TF-target gene and miRNA-target gene interactions (with data sourced from TarBase and miRTarBase). It works by mapping significant genes (such as those found differentially expressed in an RNA-seq experiment) to the corresponding molecular interaction database, and the resulting network can be exported to a Cytoscape-friendly input format. ORegAnno contains information about regulatory regions, TF binding sites, RNA binding sites, regulatory variants, haplotypes, and other regulatory elements for 18 species. Finally, rSNPBase contain information about SNPs on regulatory networks facilitating genetic studies, especially QTL studies.

In the context of *cis*-regulatory elements, this step of GRN inference can be performed either by ChIP-chip experimental approaches or using computational tools from the MEME suite (Bailey et al., 2009), which is a user-friendly web tool (**Table 3**).

2.3. Protein-protein interaction networks provide an integrated view of the proteome's organization and interactions

Proteins are intrinsically involved in every aspect of cellular bioprocesses. Simplistically, they do so by interacting with other proteins and other biocomponents and the resulting interactions may be strong or transient depending on the biological mechanisms at hand. Thus, the analysis of protein-protein interactions is a valuable way to study protein complexes, protein function annotation, and states of health and disease (Barabási et al., 2011; Snider et al., 2015).

To begin understanding the emergent characteristics of PPI one has to retrieve interaction data, which can be obtained from high-throughput techniques, interaction databases, or interaction prediction algorithms. The yeast two-hybrid (Y2H) experimental approach verifies the binary interactions between proteins by fusing them to separate *Gal4* transcription factor DNA binding and activating domains (BD and AD, respectively). The principle of the

technique relies on the interaction of a protein fused to BD, called *bait*, to the protein fused to AD, called *prey*. If *bait* and *prey* proteins interact, so do BD and AD, restoring the transcription factor activity which is reported in the assay. The Y2H is scalable and can be used to test protein interaction of many proteins in parallel with some automatization (Fields and Song, 1989).

Along with Y2H, the affinity precipitation coupled to mass spectrometry (AP-MS) yields high-throughput interaction data. Affinity purification methods use the specificity of antibody-epitope interaction to co-purify tightly interacting proteins (Bauer and Kuster, 2003). Coupling the purification phase to an identification step using MS provides means to massively generate interaction data. More PPI data can be retrieved from primary databases that store interaction information from experimental data or computational methods for interaction prediction that may involve protein sequence comparison, interologs comparison, protein surface docking, or evolutionary information using co-mutation profiles (Liu et al., 2008; Schoenrock et al., 2017; Wiles et al., 2010).

The nodes in a PPI network are proteins, and an edge is formed between a protein pair when there is evidence of interaction between them (**Table 1**). Interaction evidence may be accompanied by a score or by the qualification of that evidence, which can be set as an edge attribute to weight the support for that interaction. Usually, scores are calculated to assess the confidence in the interaction, *i.e.,* whether the interaction is confirmed by experimental and/or computational methods. The edges in a PPI network are usually undirected, but depending on the specific objective of the reconstruction it could also be set as a directed network (Vinayagam et al., 2011, 2016).

User-friendly tools for constructing protein-protein interaction networks

Many online resources of PPI data are available from different experimental or computational methods and for diverse organisms in varying conditions. The webpage Pathguide[1] presents a comprehensive list of metabolic pathways and molecular interaction resources available online and indicating if the resources are free to access, whether they follow a systems biology standard for information description and if they are still available. On the protein-protein interaction section of Pathguide there are 320 listed databases, from which 246 are still online and accessible. On **Table 4** we have listed some general protein-protein database resources. The databases listed are either free or available through academic licensing, with

the exception of STRING, which is free to use online, but in order to download the whole database a license must be purchased. The databases are classified as primary, when they gather experimental or literature-based knowledge, or secondary when they gather predicted protein interactions or reflect only a portion of the information available from primary databases (usually performing secondary analyses therein). The DIP database (Salwinski et al., 2004; Xenarios et al., 2000) has experimental interaction information that is curated automatically and manually giving the data high accuracy. STRING, which was briefly presented in the Introduction, is a database that provides experimental and/or predicted protein interaction data for over 5,000 organisms. From those listed in **Table 4** it is the only database that needs a license purchase in order to have full access to offline data, while online use of the database is free. The IntAct database (Hermjakob et al., 2004; Kerrien et al., 2012) is open-source and maintained by the European Bioinformatics Institute (EBI), gathering experimental protein-protein and protein-compound interaction data. With both protein and genetic interaction data from experimental studies, BIOGRID is a freely available primary database (Chatr-Aryamontri et al., 2017; Stark et al., 2006). It is an excellent source of curated experimental data for many model organisms and especially valuable for budding and fission yeasts. The MINT database (Chatr-Aryamontri et al., 2008) provides interaction data derived from the literature and is freely accessible. The I2D database (Brown and Jurisica, 2005, 2007) is available online and provides data for human protein-protein interactions which it imported from primary databases. It can also derive protein-protein interaction data for other model organisms if they can be mapped to human data. The Center for Cancer Systems Biology provides a primary interaction database named CCSB Interactome Database (http://interactome.dfci.harvard.edu/). The CCSB Interactome Database has experimental binary interaction data for model organisms which can be downloaded and searched freely. APID is a secondary database (Alonso-López et al., 2019) which gathers information from many primary databases, including the Protein Data Bank where protein structures are defined with interacting proteins. As an online web-tool, APID provides the possibility to select interaction properties and interactive mapping of the functional environment of proteins. HuRI, a derivation of the CCSB Interactome Database, is a database with binary protein-protein interactions for the human proteome and has three proteome scale protein-protein network reconstructions for the human genome available. Finally, the IID (Kotlyar et al., 2016) database provides tissue-specific interaction data for model organisms and human, harboring both experimental and predicted interactions.

To analyze interaction data, as for the other two previously discussed network approaches, programmable and GUI options are available. For more advanced users with a programming background, tools such as iGraph and NetworkX allow for automation and processing of large-scale datasets (Csardi and Nepusz, 2006; Hagberg et al., 2013), but user-friendly alternatives also exist, which are compiled in **Table 5**. The first step towards constructing a protein interaction network (PIN) is to get interaction data for proteins of interest. This can be done either by experimentation, as briefly described earlier, and/or by retrieving interaction data from the primary and secondary interaction databases described earlier. Interaction data can be directly downloaded or indirectly retrieved using programs or plugins, as is the case for Cytoscape. On the *Interaction database* category in **Table 5** we list Cytoscape apps that can be used to interrogate and retrieve interaction data from various databases. Bisogenet searches for molecular interaction data from an in-house database, SysBiomics, which integrates data from other interaction databases such as DIP, BIOGRID, BIND, MINT and IntAct. The searches can be filtered to narrow the interaction space, and protein annotations are retrieved from NCBI, Uniprot, KEGG, and GO. The Bisogenet app also includes PIN analysis tools. CyPath2 searches for interaction data from the Pathway Commons integrated BioPAX pathway database. PSICQUIC is a built-in feature of Cytoscape that harbors over 10 million binary interactions from 22 active data providers. The list of active providers of interaction data for PSICQUIC can be seen at the PSICQUIC Registry page[2]. StringApp imports protein-protein interaction data from STRING with a user provided protein list (or gene, compound, or disease list). Once imported, a matching network of interactions is disclosed, and functional enrichment analysis can be subsequently performed. The previously cited NetworkAnalyst is an online tool for multi-*omics* analysis, also allowing PPI visualization and analysis. It can take a network in standard format, render visualizations and perform network analysis, also receiving a gene list as input to construct an interaction network. Another online option is PINA (Protein Interaction Network Analysis platform), which generates PINs from a single protein, a list of proteins, a list of protein pairs or two lists of proteins. Networks generated by PINA can be modified with custom data or with different information from other public interaction databases. Lastly, DeDal is a Cytoscape app that embeds data information into the layout of the network, which can facilitate the user in data interpretation (**Table 5**).

For PPI network analysis, besides the previously described online resources, Cytoscape apps can be used. Apps with the *PPI-Network* tag (**Table 5**) can be applied to study the resulting

network. CyNetSVM, specifically geared towards identification of cancer biomarkers, takes as input PINs and applies artificial intelligence techniques with gene expression data to aid in the prediction of clinical outcome. CytoGEDEVO is a Cytoscape app that is capable of aligning networks, especially PINs, which can be used to study the evolution and conservation of proteins interactions. A different approach on comparison of PPIs is used by the online application INTERSPIA (INTER-Species Protein Interaction Analysis), which is freely available. INSTERSPIA can identify interacting proteins in a user-specified list and disclose similar interaction patterns across multiple species. PE-measure, another Cytoscape app, can be used to confirm protein interactions in a network based on its structure, also helping users to identify spurious interactions. Further analysis in PPI networks can be achieved using other tools in Cytoscape. PEPPER, for instance, identifies protein complexes or pathways that are highly condensed using a gene set list as input, helping to integrate information such as protein connections with proteins on the gene set list that are involved in a particular phenotype change, *e.g.,* disease, by finding functional modules. PINBPA is another app that aids in module discovery and is especially suited to integrate GWAS data into protein-protein networks, which can help identify enriched sub-networks and prioritize relevant genes. In Section 2.4 we return to the identification of modules in networks in general using algorithms that rely only on the network topology. Finally, PathLinker, a Cytoscape app, can infer signaling networks from protein-protein interaction networks by computing short paths in a PIN between receptor proteins, as source nodes, and target proteins, as transcription factors.

2.4. A primer on network analysis and visualization

Once a network of interest is attained, downstream analyses are warranted to extract relevant information and gain knowledge from the reconstruction. These analyses can be broadly divided into *knowledge extraction* and *visualization* steps. There are many methods to evaluate a network and leverage knowledge to help guide interpretation, and usually begins by exploring local and global interactions within the network. Metrics such as modularity, degree distribution, and other centrality measures are commonly applied to assist in the identification of important or influential nodes in a network (Barabási, 2016; Freeman, 1978; Jeong et al., 2001) (see **Box 1**). Cytoscape has the built-in plugin *NetworkAnalyzer* (Assenov et al., 2008) that computes many centrality metrics, and these can be extended by the Centiscape plugin, which implements ten centrality indexes (Scardoni et al., 2009). Gephi also provides built-in methods to calculate betweenness, eigenvector, and closeness centrality

measures, while bridging centrality can be calculated via a third-party plug-in (Bastian et al., 2009). Different centrality methods will usually arrive at distinct rankings of important nodes, which is not unexpected since in order to establish importance each method takes into account different aspects of the data. Betweenness centrality, for instance, emphasizes the importance of a node by considering its contribution in allowing information to pass from one part of the network to the other (thus, a global measure of centrality), while degree centrality simply counts the number of connections between a node and its direct neighbors (thus, a local measure of centrality). For some applications, a combination of centrality metrics may be more appropriate, as has been suggested for metabolic network analysis (Rio et al., 2009). In **Box 1** we present a comparison between selected centrality measures using a toy network, but an exhaustive evaluation is out of the scope of the current work, and efforts have been made to categorize and describe the various centrality indexes, such as the CentiServer online resource (http://www.centiserver.org) (Jalili et al., 2015), which harbors 232 measures of centrality in its last 2017 update, allowing users to input a network and calculate 55 centralities indexes in an interactive web-based application. The use of centrality measures in biological networks dates back to 2001, when Jeong *et al.* (2001) postulated the 'centrality-lethality rule' using a yeast protein interaction network, and found that the most highly connected proteins in the fungi's cellular network were those more important for its survival, establishing a connection between centrality (a graph-theoretical concept) and essentiality (a biological concept).

Biological networks usually display internal structures that can be identified as subnetworks in modularity analysis (Blondel et al., 2008), which present as densely connected regions, and the disclosed modules can be visually inspected by applying, for instance, the *qgraph* approach (Epskamp et al., 2012) (**Figure 2D**). Modularity (or Q) is used as a metric for defining the partitioning of a network and increases its value with increasing network community structure (Newman, 2006). The maximum modularity for a network is $Q = 1$, but in practice values for networks with strong community structure are typically in the range of 0.3-0.7 (Newman and Girvan, 2004). Many module detection techniques have been developed in the recent years and broadly divide into clustering, decomposition, and biclustering methods, which have been subject of recent reviews (Rahiminejad et al., 2019; Saelens et al., 2018). Another use of this approach is to infer biological functions using the guilty-by-association principle, where the role of an uncharacterized gene (or protein) can be predicted by considering the broad functions of the genes with which it clusters in a

modularity analysis. As an example, groups of co-expressed genes have a greater chance of being functionally coupled, either by participating in a common biological pathway or by a shared regulatory mechanism, such as an upstream regulator. In this way, novel hypotheses about gene function are generated which can be subsequently explored using as basis a co-expression network. This strategy has successfully led to the identification of novel schizophrenia risk genes, where a co-expression gene set enriched for protein-coding genes associated with the disease was disclosed (Pergola et al., 2017). As was the case for centrality metrics, both Gephi and Cytoscape offer modules to perform clustering analysis, and a Cytoscape example is shown in **Figure 5**. Gephi implements natively the Louvain algorithm, that finds modules by exploring the idea of increasing the network modularity in two phases: first, local modularity gains when neighboring nodes are included in the same cluster in an iterative fashion, which leads to local modularity maxima; second, by considering the disclosed modules from the first phase as communities and aggregating these communities iteratively (forming meta-communities) until attaining a new modularity maximum which cannot be increased further (Blondel et al., 2008). The efficiency of this algorithm allows its application to very large networks on the order of millions of nodes, one of the reasons why it has gained widespread adoption, with almost 9,000 citations (Blondel et al., 2008), including its application to disclose modules related to hepatic dysfunction (Soltis et al., 2017) and cancer (Ajorloo et al., 2017). Other clustering methods available in Gephi through third-party plugins are the Leiden (Traag et al., 2019) and the Girvan-Newman algorithms (Girvan and Newman, 2002). Girvan-Newman works by sequentially removing edges from the network until reaching a maximum modularity, and the nodes that remain connected in the resulting network represent the communities. It has been applied to a wealth of problems (accumulating over 11,000 citations), including to the successful recovery of communities of taxonomically-related organisms using protein sequence data as input (Andrade et al., 2011), but has the drawback of scaling cubically with the number of nodes in its worst case scenario, which limits its use to networks having not more than a few thousand nodes (Girvan and Newman, 2002; Rahiminejad et al., 2019). The Leiden method appeared more recently and claims to improve the quality of the disclosed modules compared to Louvain's method, as well as address some of its shortcomings (Traag et al., 2019). Other clustering methods are available through Cytoscape packages such as *clusterMaker* (Morris et al., 2011) and *CytoCluster* (Li et al., 2017b), with the latter implementing six clustering methods including OH-PIN. In contrast to the previous algorithms that only detect modules containing non-

overlapping elements, OH-PIN discloses overlapping clusters typical of many biological networks, such as enzymes that catalyze reactions across multiple pathways.

Once a network is constructed and analyzed from a topological standpoint using the previous approaches, several layout algorithms can be employed to generate visualizations of the network. While different visualization strategies do not alter the connectivity patterns between nodes, they aid during the identification of influential nodes and communities, while also allowing the organization of the network according to specific properties it may present, such as an underlying node hierarchy. Many layout algorithms are constrained by network size and can perform poorly (consuming extensive memory and CPU) when applied to the ordering of very large networks. Both Gephi (Bastian et al., 2009) and Cytoscape (Shannon et al., 2003) have a plethora of built-in visualization algorithms. In order to arrive at a suitable and pleasant network visualization a number of trial-and-error is involved, not only by qualitatively selecting layout algorithms (which can be coupled in sequence), but also by experimenting with different parameterizations schemes. Force-based algorithms are widely used to arrange networks and follow the general rule that linked nodes attract each other and non-linked nodes are mutually repelled, with inspiration from mechanical forces such as tension and compression acting through a spring, temperature gradients, or even electromagnetic forces. These methods rely only on the topology of the graph in order to arrange the nodes. Consequently, networks laid out according to force-directed strategies usually present similar edge lengths which have a low number of crossings, resulting in an aesthetically pleasing visualization. In Cytoscape, force-directed-based algorithms include the compound spring embedder and prefuse force-directed spring layout, while Gephi implements ForceAtlas2, Fruchterman-Reingold, Yifan-Hu, and OpenOrd. OpenOrd is particularly suitable for large graphs, scaling well for networks over 1 million nodes, and can be followed by the Yifan-Hu layout in order to produce appealing visualizations in such large networks (Pavlopoulos et al., 2017). Both Gephi and Cytoscape can expand their repertoire of layout methods using third-party plugins, such as the proprietary yFiles plugin for Cytoscape which offers nine options for network layout, many of which are multi-purpose such as the force-directed organic (which works well for large graphs) and orthogonal layouts (best applicable to medium-sized networks, routing edges orthogonally), as well as the hierarchic (useful for portraying precedence relationships) and circular layouts (producing star and ring topologies that are useful for visualization of regulatory relationships).

3. *Networks, networks everywhere*: health and disease from a global standpoint

Networks are now widely employed to help make sense of high-throughput *omics* data. **Figure 6** shows that usage of the networks methods that were covered in this Review are on the rise in the scientific literature. Particularly in the last 5 years, there has been a steep increase in their adoption, especially for co-expression networks, which can be partly due to the falling of sequencing costs, but also to the recent availability of some of the more user-friendly tools that were put available and reviewed herein.

Integrative approaches are particularly suitable for the study of diseases, as they are hardly the effect of single perturbations. These networks allow the identification of associations between the measured components as well as identifying communities (or modules) that could mediate a link between normal and diseased states, including regulatory interactions. Applications of correlation networks include hub genes identification in several diseases such as cancer (Oh et al., 2015), chronic fatigue syndrome (Presson et al., 2008), diabetes (Keller et al., 2008), and in the multivariate disease autism (Voineagu et al., 2011). The use of networks in the context of the neglected tropical disease leishmaniasis was also recently reviewed (Veras et al., 2018). Also performed were the stratification of breast cancer subtypes using human plasma metabolomics (Fan et al., 2016), the study of extracellular proteins in serum to disclose information on human disease states (Emilsson et al., 2018), and the evaluation of coordinated expression patterns in different brain regions in Alzheimer's disease (Wang et al., 2016a). These many studies revealed important pathways and networks of interconnected bioelements that associate with health and disease phenotypes. Co-expression and correlation networks were also used to understanding the immune response of humans to vaccination, disclosing vaccine-induced transcriptional signatures that correlated to protection (Li et al., 2017c; Nakaya et al., 2015), and have also been derived from multi-*omics* data to the understanding and tackling of disease complications from diabetes-tuberculosis comorbidity, where a correlation network constructed from whole-blood gene expression and plasma cytokine measurements was obtained (Prada-Medina et al., 2017).

Finally, disease-disease association uses the information of disease-modules in order to identify common nodes (proteins, genes, metabolites) between diseases which can help pinpoint disease comorbidity or predisposition between conditions. This approach can potentially accelerate drug design since drugs that target interactions that are common between conditions could have a better treatment impact (Barabási et al., 2011). These

methods were widely employed to construct disease-disease and gene-disease networks (Dong et al., 2018; Li et al., 2017a; Liu et al., 2018; Serão et al., 2011; Wiredja et al., 2017; Zhang et al., 2018).

While co-expression and protein-protein interaction networks are tightly related, they are both under the control of regulatory elements, thus the importance of GRNs. Environmental stimuli, pathogen exposure and other disease statuses can trigger a myriad of responses in a cell, including the cascade signals that are recognized by transcription factors, which in response modulate gene expression. Due to the specificity of GRNs for the conditions of interest, there are multiple GRNs that were generated from specific conditions, such as tissues, environments, pathologies, and the combination of these factors (Emmert-Streib et al., 2014; Guan et al., 2012). This availability of networks from specific conditions can be used to support other studies with similar conditions or used to improve GRNs for other species. In this context, GRNs can be used in health as maps and biomarkers to characterize genetic perturbations associated to rare hereditary variants such as SNPs in the regulatory region of a disease-related gene of interest (Guo and Wang, 2018).

4. Conclusions

A variety of tools are available to support the construction of biological networks from *omics* data. Although user-friendliness is usually not a top priority for developers, it can be readily attained with the help of excellent frameworks such as Cytoscape, for which a multitude of plugins are available that permits greatly expanding the capacities of the software beyond its original scope. Also, webserver versions of hitherto command-line only software are increasingly being published. We expect that user empowerment through the breaking of barriers imposed by programming language requirements will allow further adoption of network strategies and accelerate the extraction of knowledge and insights from biological data.

Box 1 - Key concepts applied to biological networks

Biological networks are composed of nodes that can represent different bioentities and have different biological importance for a given network. Regardless of the network size, shared commonalities exist between different biological networks, which allow their comparison.

The concepts below describe some characteristics of biological networks and different metrics for topological evaluation of nodes, allowing for prioritization of important elements in the network.

Scale-free. A network is considered scale-free when its degree distribution follows a power law. Thus, it is characterized by the presence of many small-degree nodes together with a few highly connected nodes (or hubs), forming an inhomogeneous network. Many biological networks exhibit the scale-free property, including protein interaction and gene co-expression networks.

Small-world. When networks exhibit a low number of node intermediates separating any two nodes in the network (*ie.*, low average distance), it is considered a small-world network.

Modularity. Biological networks tend to form modules, or clusters of highly connected nodes (**Figure box A**). Modularity takes values between -1 and 1 and reflects the link density within a module as compared to links between modules. In biological networks, nodes with similar functions have a bias to form functional modules.

Hubs. The most highly linked nodes in a network are called hub nodes, which play an important role in defining network scale-freeness. The term is also used to refer to nodes that display high centrality as measured using a relevant metric (see below).

Shortest (or geodesic) path. A shortest path is the minimum series of edges that should be traversed to connect two nodes in a network. In a weighted graph, it is the path lending to the minimum sum of edge weights between a node pair.

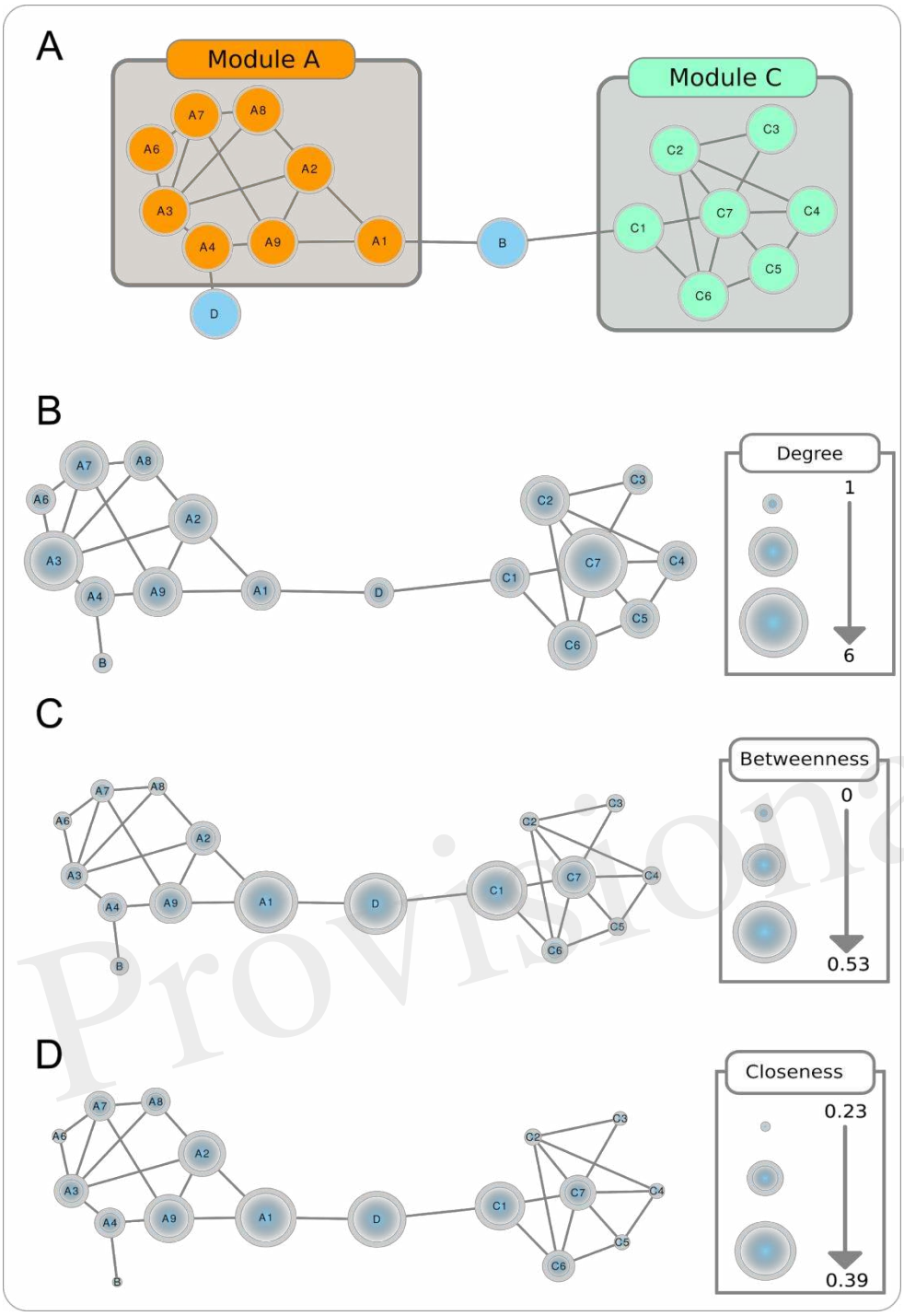
Node centrality metrics

Each component of a network presents topological characteristics that can be translated into biological knowledge and help establish the identification of relevant nodes:

Node degree. Refers to the number of nodes directly connected to a specific node, and is obtained by counting the number of interactions that a specific node has with other nodes in the network (**Figure box B**). When the network is directed, this is separated into out-degree (the number of outgoing links from a node) and in-degree (the number of ingoing links in a node). The higher is the degree of a node, the higher will be the probability that it is a hub. Nodes with high degree centrality have more influence on the structure and functionality of a network than nodes with a low degree.

Betweenness centrality. Measures the importance of a node to the connection of different parts of a network (**Figure box C**). The betweenness centrality for a node is the proportion, among all shortest paths, of those that use the given node as intermediate. Nodes with these characteristics are usually referred as bottlenecks and can also be considered hubs.

Closeness centrality. Measures how close a node is to all the other nodes in the network (**Figure box D**). It is calculated by the reciprocal sum of all shortest paths to all other nodes of the network. The higher the closeness centrality for a node, the closer is the relationship with the remaining nodes in the network.



Topological properties of a toy network. The modular aspect of the network is apparent in **A**, with two modules (or partitions) shown. The size of the nodes in **B-D** are proportional to, respectively, the node degree, betweenness centrality, and closeness centrality.

705

Footnotes

[1]http://www.pathguide.org; the webpage is maintained by Dr. Gary Bader at the University of Toronto.

[2]Available at <http://www.ebi.ac.uk/Tools/webservices/psicquic/registry/registry?action=STATUS>.

Authors' Contributions

PR conceived the review scope and outline. AQ, KF, LA, NL, and PR wrote the review. PR edited the final version with support from the other authors. All authors read and approved the final version.

Conflict of Interest Statement

The authors declare that there are no commercial or non-commercial conflicts of interest to disclose.

Funding

NL received financial support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil [Universal 28/2018; grant protocol 427183/2018-9]. LA received a postdoctoral fellowship from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). AQ acknowledges funding from Fundação Oswaldo Cruz (INOVA - Process VPPIS-001-FIO-18-45). Publication fees were defrayed by Fundação Oswaldo Cruz. The funders had no role in study design, analysis, decision to publish, or preparation of the manuscript.

Data availability statement

No datasets were generated or analyzed for this study.

References

Aittokallio, T., and Schwikowski, B. (2006). Graph-based methods for analyzing networks in cell biology. *Brief. Bioinform.* 7, 243–255.

Ajorloo, F., Vaezi, M., Saadat, A., Safaee, S. R., Gharib, B., Ghanei, M., et al. (2017). A systems medicine approach for finding target proteins affecting treatment outcomes in patients with non-Hodgkin lymphoma. *PLoS One* 12, e0183969.

Alonso-López, D., Campos-Laborie, F. J., Gutiérrez, M. A., Lambourne, L., Calderwood, M. A., Vidal, M., et al. (2019). APID database: redefining protein–protein interaction experimental evidences and binary interactomes. *Database* 2019. doi:10.1093/database/baz005.

Andrade, R. F. S., Rocha-Neto, I. C., Santos, L. B. L., de Santana, C. N., Diniz, M. V. C., Lobão, T. P., et al. (2011). Detecting network communities: an application to phylogenetic analysis. *PLoS Comput. Biol.* 7, e1001131.

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S. L., Ceol, A., Chautard, E., et al. (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat. Methods* 8, 528–529.

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics* 24, 282–284.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–8.

Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68.

Barabási, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113.

Bastian, M., Heymann, S., Jacomy, M., and Others (2009). Gephi: an open source software for exploring and manipulating networks. *Icwsm* 8, 361–362.

Bauer, A., and Kuster, B. (2003). Affinity purification-mass spectrometry: Powerful tools for the characterization of protein complexes. *Eur. J. Biochem.* 270, 570–578.

Berlin, R., Gruen, R., and Best, J. (2017). Systems Medicine—Complexity Within, Simplicity Without. *Journal of Healthcare Informatics Research* 1, 119–137.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008, P10008.

Brown, K. R., and Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics* 21, 2076–2082.

Brown, K. R., and Jurisica, I. (2007). Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.* 8, R95.

Brown, K. R., Otasek, D., Ali, M., McGuffin, M. J., Xie, W., Devani, B., et al. (2009). NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. *Bioinformatics* 25, 3327–3329.

Cardozo, L. E., Russo, P. S. T., Gomes-Correia, B., Araujo-Pereira, M., Sepúlveda-Hermosilla, G., Maracaja-Coutinho, V., et al. (2019). webCEMiTool: Co-expression Modular Analysis Made Easy. *Front. Genet.* 10, 146.

Carpenter, A. E., and Sabatini, D. M. (2004). Systematic genome-wide screens of gene function. *Nat. Rev. Genet.* 5, 11–22.

Chai, L., Loh, S. K., Low, S. T., Mohamad, M. S., Deris, S., Zakaria, Z. (2014). A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine* 48, 55–65.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., et al. (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* 45, D369–D379.

Chatr-Aryamontri, A., Zanzoni, A., Ceol, A., and Cesareni, G. (2008). Searching the protein interaction space through the MINT database. *Methods Mol. Biol.* 484, 305–317.

Cottret, L., and Jourdan, F. (2010). Graph methods for the investigation of metabolic networks in parasitology. *Parasitology* 137, 1393–1407.

Cowley, M. J., Pinese, M., Kassahn, K. S., Waddell, N., Pearson, J. V., Grimmond, S. M., et al. (2012). PINA v2.0: mining interactome modules. *Nucleic Acids Res.* 40, D862–5.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems* 1695, 1–9.

Czerwinska, U., Calzone, L., Barillot, E., and Zinovyev, A. (2015). DeDaL: Cytoscape 3 app for producing and morphing data-driven and structure-driven network layouts. *BMC Syst. Biol.* 9, 46.

de Matos Simoes, R., Dehmer, M., and Emmert-Streib, F. (2013). B-cell lymphoma gene regulatory networks: biological consistency among inference methods. *Front. Genet.* 4, 281.

de Siqueira Santos, S., Takahashi, D. Y., Nakata, A., and Fujita, A. (2014). A comparative study of statistical methods used to identify dependencies between gene expression signals. *Brief. Bioinform.* 15, 906–918.

Doncheva, N. T., Morris, J. H., Gorodkin, J., and Jensen, L. J. (2019). Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *J. Proteome Res.* 18, 623–632.

Dong, H., Zhang, S., Wei, Y., Liu, C., Wang, N., Zhang, P., et al. (2018). Bioinformatic analysis of differential expression and core GENEs in breast cancer. *Int. J. Clin. Exp. Pathol.* 11, 1146–1156.

Dutta, N. K., Bandyopadhyay, N., Veeramani, B., Lamichhane, G., Karakousis, P. C., and Bader, J. S. (2014). Systems Biology-Based Identification of Mycobacterium tuberculosis Persistence Genes in Mouse Lungs. *mBio* 5. doi:10.1128/mbio.01066-13.

Emilsson, V., Ilkov, M., Lamb, J. R., Finkel, N., Gudmundsson, E. F., Pitts, R., et al. (2018). Co-regulatory networks of human serum proteins link genetics to disease. *Science* 361, 769–773.

Emmert-Streib, F., de Matos Simoes, R., Mullan, P., Haibe-Kains, B., and Dehmer, M. (2014). The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. *Front. Genet.* 5, 15.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., Borsboom, D., and Others (2012). qgraph: Network visualizations of relationships in psychometric data. *J. Stat. Softw.* 48, 1–18.

Fan, Y., Zhou, X., Xia, T.-S., Chen, Z., Li, J., Liu, Q., et al. (2016). Human plasma metabolomics for identifying differential metabolites and predicting molecular subtypes of breast cancer. *Oncotarget* 7, 9925–9938.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* 340, 245–246.

Floratos, A., Smith, K., Ji, Z., Watkinson, J., and Califano, A. (2010). geWorkbench: an open source platform for integrative genomics. *Bioinformatics* 26, 1779–1780.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Soc. Networks* 1, 215–239.

Fronczuk, M., Raftery, A. E., and Yeung, K. Y. (2015). CyNetworkBMA: a Cytoscape app for inferring gene regulatory networks. *Source Code Biol. Med.* 10, 11.

Gallo, C. A., Carballido, J. A., and Ponzoni, I. (2011). Discovering time-lagged rules from microarray data using gene profile classifiers. *BMC Bioinformatics* 12, 123.

Gil, D. P., Law, J. N., and Murali, T. M. (2017). The PathLinker app: Connect the dots in protein interaction networks. *F1000Res.* 6, 58.

Girvan, M., and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.* 99, 7821–7826.

Guan, Y., Gorenshteyn, D., Burmeister, M., Wong, A. K., Schimenti, J. C., Handel, M. A., et al. (2012). Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLoS Comput. Biol.* 8, e1002694.

Guitart-Pla, O., Kustagi, M., Rügheimer, F., Califano, A., and Schwikowski, B. (2015). The Cyni framework for network inference in Cytoscape. *Bioinformatics* 31, 1499–1501.

Guo, L., and Wang, J. (2018). rSNPBase 3.0: an updated database of SNP-related regulatory elements, element-gene pairs and SNP-based gene regulatory networks. *Nucleic Acids Res.* 46, D1111–D1116.

Hache, H., Lehrach, H., and Herwig, R. (2009). Reverse engineering of gene regulatory networks: a comparative study. *EURASIP J. Bioinform. Syst. Biol.*, 617281.

Hagberg, A., Schult, D., Swart, P., Conway, D., Séguin-Charbonneau, L., Ellison, C., et al. (2013). Networkx. High productivity software for complex networks. *Webová strá nka https://networkx. lanl. gov/wiki*.

Han, H., Cho, J.-W., Lee, S., Yun, A., Kim, H., Bae, D., et al. (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* 46, D380–D386.

Hardin, J., Mitani, A., Hicks, L., and VanKoten, B. (2007). A robust measure of correlation between two genes on a microarray. *BMC Bioinformatics* 8, 220.

Hartemink, A. J. (2005). Reverse engineering gene regulatory networks. *Nat. Biotechnol.* 23, 554–555.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., et al. (2004). IntAct: an open source molecular interaction database. *Nucleic Acids Res.* 32, D452–5.

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front. Genet.* 8, 84.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 5. doi:10.1371/journal.pone.0012776.

Jalili, M., Salehzadeh-Yazdi, A., Asgari, Y., Arab, S. S., Yaghmaie, M., Ghavamzadeh, A., et al. (2015). CentiServer: A Comprehensive Resource, Web-Based Application and R Package for Centrality Analysis. *PLoS One* 10, e0143111.

Janky, R. 's, Verfaillie, A., Imrichová, H., Van de Sande, B., Standaert, L., Christiaens, V., et al. (2014). iRegulon: from a gene list to a gene regulatory network using large motif and track collections. *PLoS Comput. Biol.* 10, e1003731.

Jeong, H., Mason, S. P., -L. Barabási, A., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41–42. doi:10.1038/35075138.

Jha, A. K., Huang, S. C.-C., Sergushichev, A., Lampropoulou, V., Ivanova, Y., Loginicheva, E., et al. (2015). Network integration of parallel metabolic and transcriptional data reveals metabolic modules that regulate macrophage polarization. *Immunity* 42, 419–430.

Kaderali, L., Radde, N. (2008). Inferring Gene Regulatory Networks from Expression Data. *Computational Intelligence in Bioinformatics* 94, 33–74.

Keller, M. P., Choi, Y., Wang, P., Davis, D. B., Rabaglia, M. E., Oler, A. T., et al. (2008). A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Res.* 18, 706–716.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., et al. (2012).

The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 40, D841–6.

Klein, C. C., Cottret, L., Kielbassa, J., Charles, H., Gautier, C., Ribeiro de Vasconcelos, A. T., et al. (2012). Exploration of the core metabolism of symbiotic bacteria. *BMC Genomics* 13, 438.

Kotlyar, M., Pastrello, C., Sheahan, N., and Jurisica, I. (2016). Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res.* 44, D536–41.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–7.

Kwon, D., Lee, D., Kim, J., Lee, J., Sim, M., and Kim, J. (2018). INTERSPIA: a web application for exploring the dynamics of protein-protein interactions among multiple species. *Nucleic Acids Res.* 46, W89–W94.

Lacroix, V., Cottret, L., Thebault, P., and -F. Sagot, M. (2008). An Introduction to Metabolic Networks and Their Structural Analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 5, 594–617. doi:10.1109/tcbb.2008.79.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559.

Langfelder, P., and Horvath, S. (2012). FastRFunctions for Robust Correlations and Hierarchical Clustering. *J. Stat. Softw.* 46. doi:10.18637/jss.v046.i11.

Lee, T. I., Johnstone, S. E., and Young, R. A. (2006). Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat. Protoc.* 1, 729–748.

Lesurf, R., Cotto, K. C., Wang, G., Griffith, M., Kasaian, K., Jones, S. J. M., et al. (2016). ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Res.* 44, D126–32.

Li, D.-Y., Chen, W.-J., Luo, L., Wang, Y.-K., Shang, J., Zhang, Y., et al. (2017a). Prospective lncRNA-miRNA-mRNA regulatory network of long non-coding RNA LINC00968 in non-small cell lung cancer A549 cells: A miRNA microarray and bioinformatics investigation. *International Journal of Molecular Medicine*. doi:10.3892/ijmm.2017.3187.

Li, M., Li, D., Tang, Y., Wu, F., and Wang, J. (2017b). CytoCluster: A Cytoscape Plugin for Cluster Analysis and Visualization of Biological Networks. *International Journal of Molecular Sciences* 18, 1880. doi:10.3390/ijms18091880.

Li, S., Sullivan, N. L., Rouphael, N., Yu, T., Banton, S., Maddur, M. S., et al. (2017c). Metabolic Phenotypes of Response to Vaccination in Humans. *Cell* 169, 862–877.e17.

Liu, D., Skomorovska, Y., Song, J., Bowler, E., Harris, R., Ravasz, M., et al. (2018). ELF3 is an antagonist of oncogenic-signaling-induced expression of EMT-TF ZEB1. *Cancer Biol. Ther.*, 1–11.

Liu, S., Gao, Y., and Vakser, I. A. (2008). Dockground protein–protein docking decoy set. *Bioinformatics* 24, 2634–2635.

Liu, Z.-P., Wu, C., Miao, H., and Wu, H. (2015). RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* 2015. doi:10.1093/database/bav095.

Luo, W., Hankenson, K. D., and Woolf, P. J. (2008). Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information. *BMC Bioinformatics* 9, 467.

Malek, M., Ibragimov, R., Albrecht, M., and Baumbach, J. (2016). CytoGEDEVO—global alignment of biological networks with Cytoscape. *Bioinformatics* 32, 1259–1261.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006a). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7 Suppl 1, S7.

Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I., and Califano, A. (2006b). Reverse engineering cellular networks. *Nat. Protoc.* 1, 662–671.

Martin, A., Ochagavia, M. E., Rabasa, L. C., Miranda, J., Fernandez-de-Cossio, J., and Bringas, R. (2010). BisoGenet: a new tool for gene network building, visualization and analysis. *BMC Bioinformatics* 11, 91.

Meyer, P. E., Lafitte, F., and Bontempi, G. (2008). minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics* 9, 461.

Modrák, M., and Vohradský, J. (2018). Genexpi: a toolset for identifying regulons and validating gene regulatory networks using time-course expression data. *BMC Bioinformatics* 19, 137.

Morris, J. H., Apeltsin, L., Newman, A. M., Baumbach, J., Wittkop, T., Su, G., et al. (2011). clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics* 12, 436.

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* 24, 69–71.

Nakaya, H. I., Hagan, T., Duraisingham, S. S., Lee, E. K., Kwissa, M., Rouphael, N., et al. (2015). Systems Analysis of Immunity to Influenza Vaccination across Multiple Years and in Diverse Populations Reveals Shared Molecular Signatures. *Immunity* 43, 1186–1198.

Naldi, A., Berenguiera, D., Fauréa, A., Lopeza, F., Thieffrya, D., Chaouiya, C. (2009). Logical modelling of regulatory networks with GINsim 2.3. *BioSystems* 97, 134–139.

Narasimhan, S., Rengaswamy, R., and Vadigepalli, R. (2009). Structural properties of gene regulatory networks: definitions and connections. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 6, 158–170.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U. S. A.* 103, 8577–8582.

Newman, M. E. J., and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 69, 026113.

Ngounou Wetie, A. G., Sokolowska, I., Woods, A. G., Roy, U., Deinhardt, K., and Darie, C. C. (2014). Protein-protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. *Cell. Mol. Life Sci.* 71, 205–228.

Oh, E.-Y., Christensen, S. M., Ghanta, S., Jeong, J. C., Bucur, O., Glass, B., et al. (2015). Extensive rewiring of epithelial-stromal co-expression networks in breast cancer. *Genome Biol.* 16, 128.

Pavlopoulos, G. A., Paez-Espino, D., Kyrpides, N. C., and Iliopoulos, I. (2017). Empirical Comparison of Visualization Tools for Larger-Scale Network Analysis. *Adv. Bioinformatics* 2017, 1278932.

Pergola, G., Di Carlo, P., D'Ambrosio, E., Gelao, B., Fazio, L., Papalino, M., et al. (2017). DRD2 co-expression network and a related polygenic index predict imaging, behavioral and clinical phenotypes linked to schizophrenia. *Translational Psychiatry* 7, e1006–e1006. doi:10.1038/tp.2016.253.

Prada-Medina, C. A., Fukutani, K. F., Pavan Kumar, N., Gil-Santana, L., Babu, S., Lichtenstein, F., et al. (2017). Systems Immunology of Diabetes-Tuberculosis Comorbidity Reveals Signatures of Disease Complications. *Sci. Rep.* 7, 1999.

Presson, A. P., Sobel, E. M., Papp, J. C., Suarez, C. J., Whistler, T., Rajeevan, M. S., et al. (2008). Integrated Weighted Gene Co-expression Network Analysis with an Application to Chronic Fatigue Syndrome. *BMC Syst. Biol.* 2, 95.

Rahiminejad, S., Maurya, M. R., and Subramaniam, S. (2019). Topological and functional comparison of community detection algorithms in biological networks. *BMC Bioinformatics* 20, 212.

Rio, G. del, del Rio, G., Koschützki, D., and Coello, G. (2009). How to identify essential genes from molecular networks? *BMC Systems Biology* 3. doi:10.1186/1752-0509-3-102.

Russo, P. S. T., Ferreira, G. R., Cardozo, L. E., Bürger, M. C., Arias-Carrasco, R., Maruyama, S. R., et al. (2018). CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinformatics* 19, 56.

Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* 9, 1090.

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., Lotia, S., et al. (2012). A travel guide to Cytoscape plugins. *Nat. Methods* 9, 1069–1076.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 32, D449–51.

Scardoni, G., Petterlini, M., and Laudanna, C. (2009). Analyzing biological network parameters with CentiScaPe. *Bioinformatics* 25, 2857–2859.

Schoenrock, A., Burnside, D., Moteshareie, H., Pitre, S., Hooshyar, M., Green, J. R., et al. (2017). Evolution of protein-protein interaction networks in yeast. *PLoS One* 12, e0171920.

Serão, N. V. L., Delfino, K. R., Southey, B. R., Beever, J. E., and Rodriguez-Zas, S. L. (2011). Cell cycle and aging, morphogenesis, and response to stimuli genes are individualized biomarkers of glioblastoma progression and survival. *BMC Med. Genomics* 4, 49.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504.

Shi, X., Banerjee, S., Chen, L., Hilakivi-Clarke, L., Clarke, R., and Xuan, J. (2017). CyNetSVM: A Cytoscape App for Cancer Biomarker Identification Using Network Constrained Support Vector Machines. *PLoS One* 12, e0170482.

Shrinet, J., Nandal, U. K., Adak, T., Bhatnagar, R. K., and Sunil, S. (2014). Inference of the oxidative stress network in Anopheles stephensi upon Plasmodium infection. *PLoS One* 9, e114461.

Snider, J., Kotlyar, M., Saraon, P., Yao, Z., Jurisica, I., and Stagljar, I. (2015). Fundamentals of protein interaction network mapping. *Mol. Syst. Biol.* 11, 848.

Soltis, A. R., Kennedy, N. J., Xin, X., Zhou, F., Ficarro, S. B., Yap, Y. S., et al. (2017). Hepatic Dysfunction Caused by Consumption of a High-Fat Diet. *Cell Rep.* 21, 3317–3328.

Song, Q., Grene, R., Heath, L. S., and Li, S. (2017). Identification of regulatory modules in genome scale transcription regulatory networks. *BMC Syst. Biol.* 11, 140.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34, D535–9.

Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* 18, S231–S240.

Stevens, A., De Leonibus, C., Hanson, D., Dowsey, A. W., Whatmore, A., Meyer, S., et al. (2014). Network analysis: a new approach to study endocrine disorders. *J. Mol. Endocrinol.* 52, R79–93.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368.

Tawfike, A. F., Romli, M., Clements, C., Abbott, G., Young, L., Schumacher, M., et al. (2019). Isolation of anticancer and anti-trypanosome secondary metabolites from the endophytic fungus Aspergillus flocculus via bioactivity guided isolation and MS based

metabolomics. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 1106-1107, 71–83.

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* 9, 5233.

Vella, D., Zoppis, I., Mauri, G., Mauri, P., and Di Silvestre, D. (2017). From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *EURASIP J. Bioinform. Syst. Biol.* 2017, 6.

Veras, P. S. T., Ramos, P. I. P., and de Menezes, J. P. B. (2018). In Search of Biomarkers for Pathogenesis and Control of Leishmaniasis by Global Analyses of -Infected Macrophages. *Front. Cell. Infect. Microbiol.* 8, 326.

Vinayagam, A., Gibson, T. E., Lee, H.-J., Yilmazel, B., Roesel, C., Hu, Y., et al. (2016). Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proc. Natl. Acad. Sci. U. S. A.* 113, 4976–4981.

Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., et al. (2011). A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.* 4, rs8.

Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Horvath, S., et al. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474, 380–384.

Walter, S. D., and Altman, D. G. (1992). Practical Statistics for Medical Research. *Biometrics* 48, 656. doi:10.2307/2532320.

Wang, L., Matsushita, T., Madireddy, L., Mousavi, P., and Baranzini, S. E. (2015). PINBPA: cytoscape app for network analysis of GWAS data. *Bioinformatics* 31, 262–264.

Wang, M., Roussos, P., McKenzie, A., Zhou, X., Kajiwara, Y., Brennand, K. J., et al. (2016a). Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. *Genome Med.* 8. doi:10.1186/s13073-016-0355-3.

Wang, P., Wang, Y., Hang, B., Zou, X., and Mao, J.-H. (2016b). A novel gene expression-based prognostic scoring system to predict survival in gastric cancer. *Oncotarget* 7, 55343–55351.

Wei, T., and Simko, V. (2017). R package "corrplot": visualization of a correlation matrix (version 0.84).'. *Retrieved from https://github. com/taiyun/corrplot*.

Wiles, A. M., Doderer, M., Ruan, J., Gu, T.-T., Ravi, D., Blackman, B., et al. (2010). Building and analyzing protein interactome networks by cross-species comparisons. *BMC Syst. Biol.* 4, 36.

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., et al. (2004). 10.1186/gb-2004-5-11-r92. *Genome Biol* 5, R92. doi:10.1186/gb-2004-5-11-r92.

Winterhalter, C., Nicolle, R., Louis, A., To, C., Radvanyi, F., and Elati, M. (2014). Pepper: cytoscape app for protein complex expansion using protein–protein interaction

networks. *Bioinformatics* 30, 3419–3420.

Wiredja, D. D., Ayati, M., Mazhar, S., Sangodkar, J., Maxwell, S., Schlatzer, D., et al. (2017). Phosphoproteomics Profiling of Nonsmall Cell Lung Cancer Cells Treated with a Novel Phosphatase Activator. *Proteomics* 17. doi:10.1002/pmic.201700214.

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mäkelä, T. P., and Hautaniemi, S. (2009). Integrated network analysis platform for protein-protein interactions. *Nat. Methods* 6, 75–77.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res.* 28, 289–291.

Zaki, N., Efimov, D., and Berengueres, J. (2013). Protein complex detection using interaction reliability assessment and weighted clustering coefficient. *BMC Bioinformatics* 14, 163.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4, Article17.

Zhang, F., Xu, W., Liu, J., Liu, X., Huo, B., Li, B., et al. (2018). Optimizing miRNA-module diagnostic biomarkers of gastric carcinoma via integrated network analysis. *PLoS One* 13, e0198445.

Zhang, W., Mao, J.-H., Zhu, W., Jain, A. K., Liu, K., Brown, J. B., et al. (2016). Centromere and kinetochore gene misexpression predicts cancer patient survival and response to radiotherapy and chemotherapy. *Nat. Commun.* 7, 12619.

Zhou, G., Soufan, O., Ewald, J., Hancock, R. E. W., Basu, N., and Xia, J. (2019). NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res.* doi:10.1093/nar/gkz240.

Bavelas, A. (1950). Communication patterns in task-oriented groups. Journal of the acoustical society of America, 22, 725-730.

Sabidussi, G. (1966). The centrality index of a graph. Psychometrika, 31, 581-603.

Figure Legends

Figure 1

A roadmap to network concepts covered in this review. Three simple six-node graphs are shown in the upper panel. These graphs can be undirected (**A**), directed (**B**) or weighted directed (**C**). In the latter, the thickness of edges reflects the weights of the interactions. Various *omics* datasets can be analyzed using the language of networks, which are discussed in the following sections (**D**). (**E**) Once a network is attained, further analyses are warranted, such as disclosing modules or communities and calculating topological metrics such as node degree and betweenness centrality (BC), covered in Section 2.4. The size of a node is proportional to its degree, while the color reflects the community structure in this illustrative example where two modules are disclosed. For selected nodes, interpretations of node BC and degree are presented.

Figure 2

Different views on assessing correlations. (**A**) Classic scatter plot with correlation curve (straight black line). (**B**) Correlation matrix plot, designed with the *corrplot* package (Wei and Simko, 2017). (**C**) Circular layout correlation network, designed with Gephi (Bastian et al., 2009). (**D**) Complex correlation network with modularity coloring, designed with *qgraph* package (Epskamp et al., 2012).

Figure 3

A correlation network constructed using Cytoscape 3.2. The network was built using a bacterial expression dataset, and nodes represent annotated genes, with edges connecting nodes if they pass a correlation threshold calculated using Spearman's rank correlation in the Cyni Toolbox. In the picture a pop-up menu with the calculated network metrics (using the NetworkAnalyzer plugin in Cytoscape) is shown. Besides the network zoom, the program also shows the whole network in the lower-right screen, as a miniature.

Figure 4

Different ways to represent gene regulatory networks. (**A**) Toy networks exemplifying bipartite and logical (Boolean) graphs. (**B**) A real example of the human gene regulatory network extracted from TRRUST database, and its graphical representation as a bipartite and a logical networks.

Figure 5

Typical network analyses performed using Cytoscape. A network of yeast protein interaction data is presented (**A**), with node size scaled with betweenness centrality, which help in straightforward identification of important nodes in this network. Nodes are colored according to its membership to a community as determined using the Girvan-Newman fast greedy algorithm implementation in the *clusterMaker* plugin (Morris et al., 2011). Colors for each community were chosen automatically using a color-generating function and a discrete mapping, with modules numbered sequentially in the left column shown in **B**, and colors (in RGB and hex formats) on the right. Properties of nodes are shown below in **C**, including some centrality measures. These can be downloaded in-whole as a table for downstream analyses. The network is arranged according to a force-directed layout algorithm.

Figure 6

Network methods on the rise. Searches in PubMed (http://ncbi.nlm.nih.gov/pubmed) were performed to identify the all-time use of co-expression networks (query: "co-expression network" OR "coexpression network"), gene regulatory networks (GRN; query: "gene regulatory network"), and protein-protein interaction networks (PPI; query: "protein-protein interaction network"). Data for 2019 is partial (up to March) and are displayed as open points.

Tables

Table 1 - Biological interpretation of nodes, edges, and edge weights for the *omics*-derived networks under study.

Type of network	**Graph representation**	**Edge directionality**	**Biological interpretation of**		
			nodes	*edges*	*edge weights*
Correlation network	Simple graph	Undirected	Genes, proteins, or metabolites	Correlation (co-expression) between a pair of biological entities, which is calculated from a measure of abundance, such as gene expression or metabolite concentration	The strength of correlation (co-expression) between the pair of nodes
Gene regulatory network	Simple or bipartite graphs	Usually directed	Genes in the simple graph; transcription factors and target-genes in the bipartite graph	A regulatory relationship	The degree of the regulatory relationship
Protein-protein interaction network	Simple graph	Usually undirected	Proteins	The direct contact (physical binding) between proteins, but can represent indirect (functional) interactions between the peptides	Usually unweighted, but can be valued to represent the support (confidence) for a given interaction

Table 2 - User-friendly computational tools for inferring correlation networks.

Tool	**Description**	**Platform**	**Reference/URL**
Cyni toolbox (Cytoscape)	Performs several correlation analyses and includes other networks inference algorithms.	Multi	http://apps.cytoscape.org/apps/cynitoolbox; (Guitart-Pla et al., 2015)
Expression Correlation app (Cytoscape)	Performs Pearson correlation analysis and network inference.	Multi	http://apps.cytoscape.org/apps/expressioncorrelation

ARACNe/Mutual Information (geWorkbench)	Creates a network based on Mutual Information.	Multi	http://wiki.c2b2.columbia.edu/workbench/index.php/Home; (Floratos et al., 2010)
webCEMiTool	Performs comprehensive modular analyses in a fully automated manner, generating co-expression networks based on the WGCNA method.	Webserver	https://cemitool.sysbio.tools/; (Cardozo et al., 2019)

Table 3 - User-friendly computational tools for inferring gene regulatory networks.

Tool	**Description**	**Platform**	**Type of data**		**Reference/URL**
			Expression	*Promoter*	
ARACNe	Creates a network based on Mutual Information	Multi	✓		http://apps.cytoscape.org/apps/aracne; (Floratos et al., 2010)
CyGenexpi	A toolset for identifying regulons and validating gene regulatory networks using time-course expression data	Multi	✓		https://apps.cytoscape.org/apps/cygenexpi; (Modrák and Vohradský, 2018)
CyNetworkBMA	Infers gene regulatory networks from expression measurements using Bayesian Model Averaging	Multi	✓		https://apps.cytoscape.org/apps/cynetworkbma; (Fronczuk et al., 2015)
GRNCOP2	Model-free combinatorial optimization algorithm to infer time-delayed gene regulatory networks from genome-wide time series datasets	Multi	✓		https://apps.cytoscape.org/apps/grncop2; (Gallo et al., 2011)
iRegulon	Allows identification of regulons using motif and track discovery in an existing network	Multi		✓	https://apps.cytoscape.org/apps/iregulon; (Janky et al., 2014)
NetworkAnalyst	Allows establishing TF-target genes and miRNAs-target genes associations.	Webserver	✓		http://www.networkanalyst.ca; (Zhou et al., 2019)
TRRUST	TFs and target genes interactions, and TFs cis-regulatory elements	Webserver	✓	✓	https://www.grnpedia.org/trrust/Network_search_form.php; (Han et al., 2018)

RegNetwork	Genic regulations by TFs and microRNAs	Webserver	✓		http://www.regnetworkweb.org/search.jsp; (Liu et al., 2015)
ORegAnno	Regulatory regions, transcription factor binding sites, etc.	Webserver		✓	http://www.oreganno.org/; (Lesurf et al., 2016)
rSNPBase	Harbors curated information on regulatory SNPs	Webserver		✓	http://rsnp.psych.ac.cn/; (Guo and Wang, 2018)
MEME	Sequence analysis tools for motifs discovery	Webserver		✓	http://meme-suite.org/ (Bailey et al., 2009)

1180

Table 4 - On-line resources for acquiring protein interaction information.

Abbreviation	**Name**	**URL**	**Availability**	**Data Source**
DIP	Database of Interacting Proteins	http://dip.doe-mbi.ucla.edu/dip/Main.cgi	Academic license	Primary
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins	http://string-db.org/	License purchase	Secondary
IntAct	IntAct Molecular Interaction Database	http://www.ebi.ac.uk/intact	Free	Primary
BioGRID	Biological General Repository for Interaction Datasets	http://www.thebiogrid.org/	Free	Primary
MINT	Molecular Interaction Database	http://mint.bio.uniroma2.it/	Free	Primary
I2D	Interologous Interaction Database	http://ophid.utoronto.ca/	Academic license	Secondary
CCSB	Center for Cancer Systems Biology Interactome Database	http://interactome.dfci.harvard.edu/	Free	Primary
APID	Agile Protein Interactomes DataServer	http://apid.dep.usal.es/	Free	Secondary
HuRI	The Human Reference Protein Interactome Mapping Project	http://interactome.baderlab.org/	Academic license	Primary
IID	Integrated Interactions Database	http://iid.ophid.utoronto.ca/iid/Search_By_Proteins/	Academic license	Primary

1185 Table 5 - User-friendly computational tools for inferring and analyzing protein interaction networks.

Tool	Description	Category	Reference/URL
Bisogenet	Retrieves interactions associated with input IDs. Sophisticated UI gives links to GO, KEGG, etc.	*Interaction database*	(Martin et al., 2010)
CyNetSVM	Developed for identification of cancer biomarkers using machine learning approaches.	*PPI-network*	(Shi et al., 2017)
CyPath2	Pathway Commons (BioPAX L3 database) web service GUI client app.	*Interaction database*	http://apps.cytoscape.org/apps/cypath2
CytoGEDEVO	Pairwise global alignment of PPI or other networks.	*PPI-network*	(Malek et al., 2016)
CytoMOBAS	Identifies and analyses disease associated and highly connected subnetworks.	*Disease-disease association* *PPI-network*	https://apps.cytoscape.org/apps/cytomobas
DeDal	Applies data dimensionality reduction methods for designing insightful network visualizations.	*PPI-network*	(Czerwinska et al., 2015)
INTERSPIA	Free online resource for protein interaction comparison between species	Not a Cytoscape app	(Kwon et al., 2018)
NetworkAnalyst	Free online resource for network construction and analysis	Not a Cytoscape app	(Zhou et al., 2019)
PathLinker	Reconstructs the interactions in a signaling pathway of interest from the receptors and TFs in a pathway, and can be broadly used to compute and analyze a network of protein interactions.	*PPI-network*	(Gil et al., 2017)
PEmeasure	Compute links weights and assess the reliability of the links in a network including PPI.	*PPI-network*	(Zaki et al., 2013)
PEPPER	Find meaningful pathways / complexes connecting a protein set members within a PPI-network using multi-objective optimization.	*Functional module detection*	(Winterhalter et al., 2014)
PINA	Free online resource capable of PIN construction, filtering, analysis, visualization and management.	Not a Cytoscape app	(Cowley et al., 2012; Wu et al., 2009)
PINBPA	Protein-interaction-network-based Pathway Analysis.	*Random walk with restart algorithm*	(Wang et al., 2015)
PSICQUIC	PSICQUIC Web Service Client for	*Interaction database*	(Aranda et al., 2011)

Universal Client	importing interactions from public databases.		
stringApp	Import and augment Cytoscape networks from STRING.	*Gene-disease association; PPI-network*	(Doncheva et al., 2019)

Figure 01.TIF

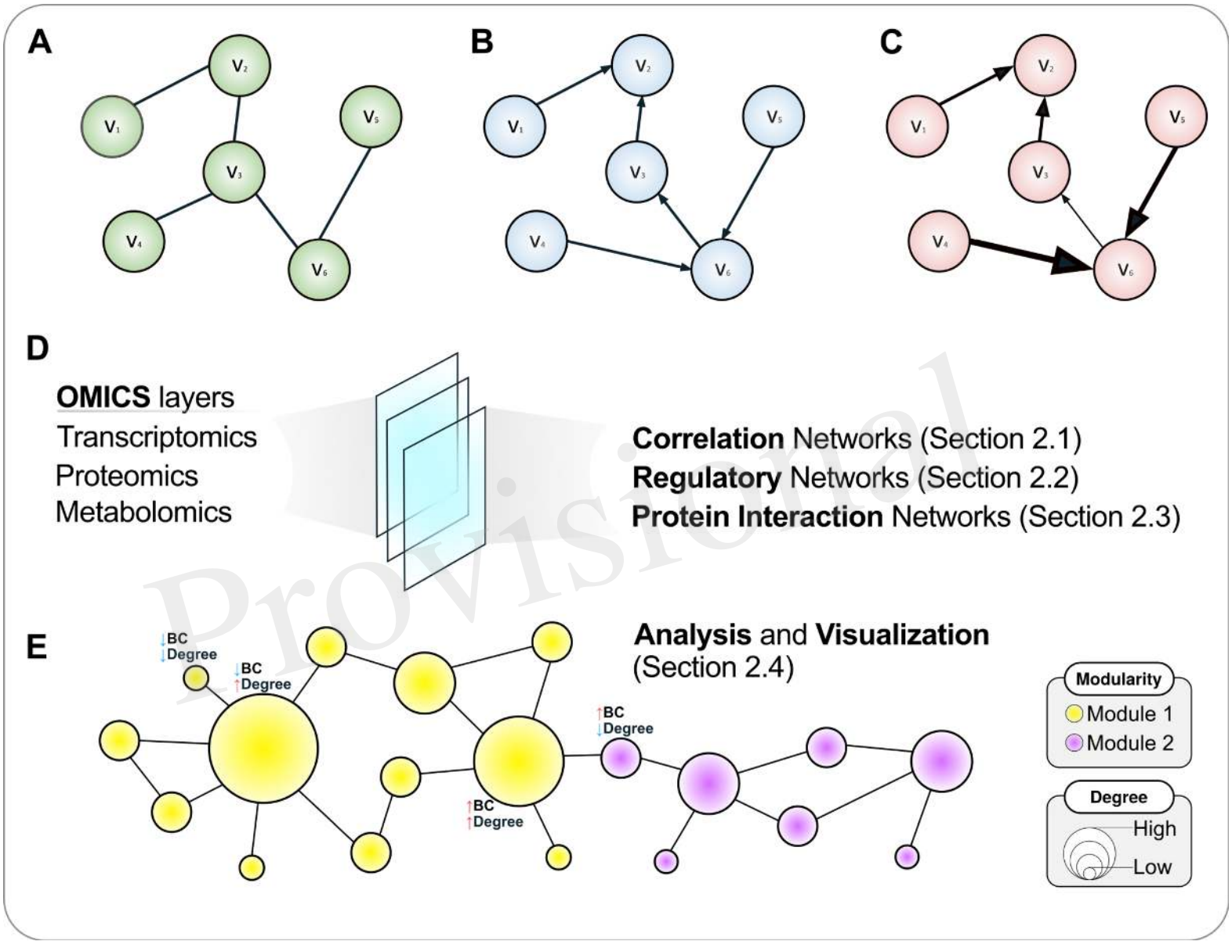


Figure 02.TIFF

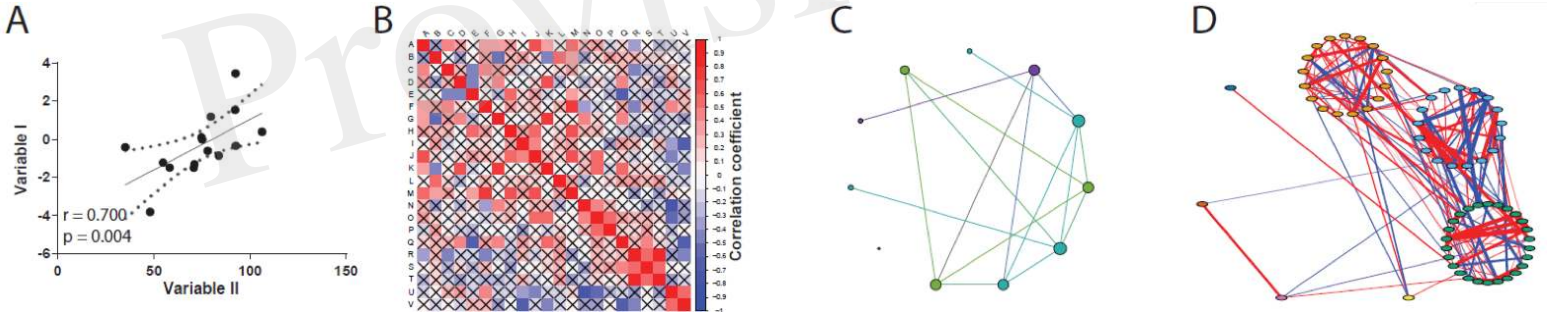


Figure 03.TIFF

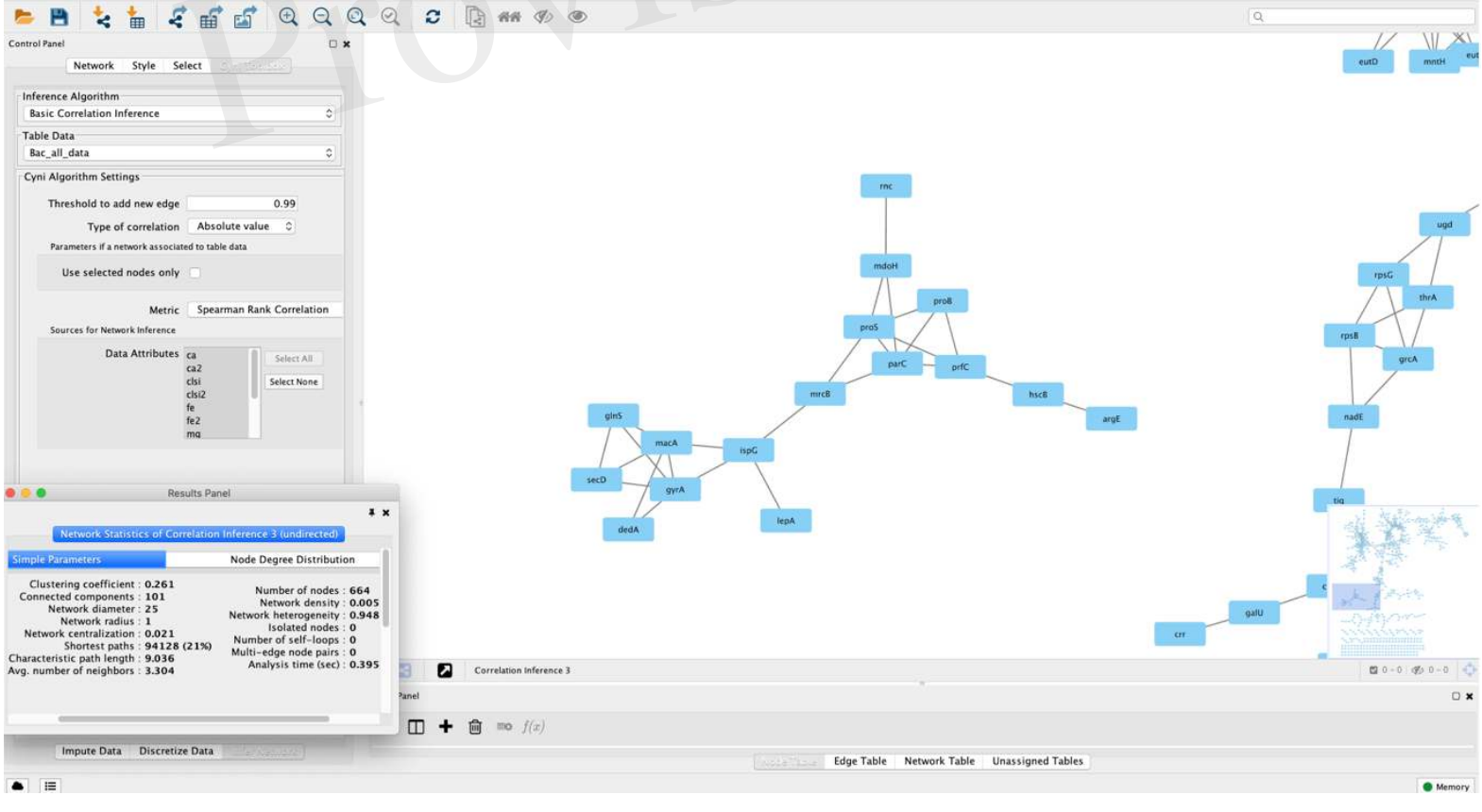


Figure 04.JPEG

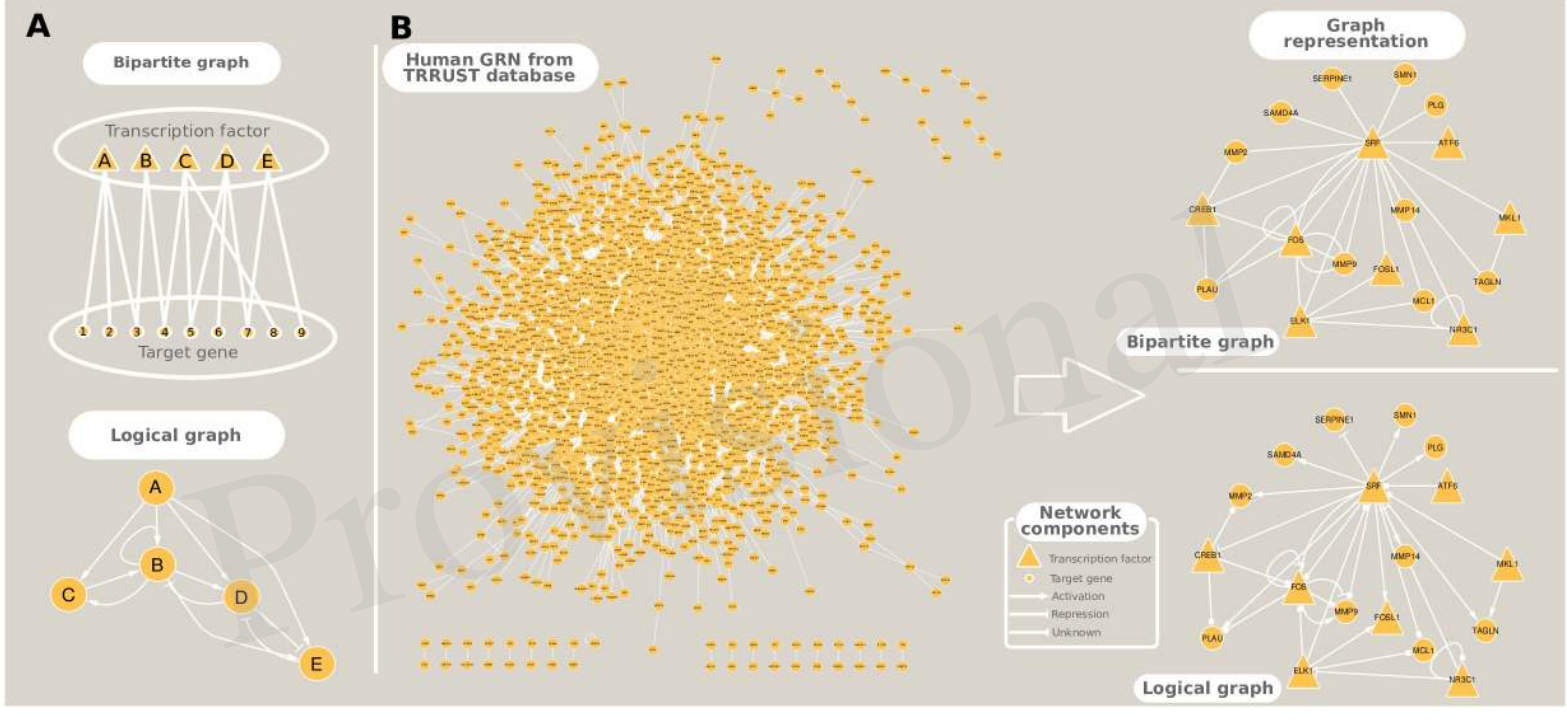


Figure 05.JPEG

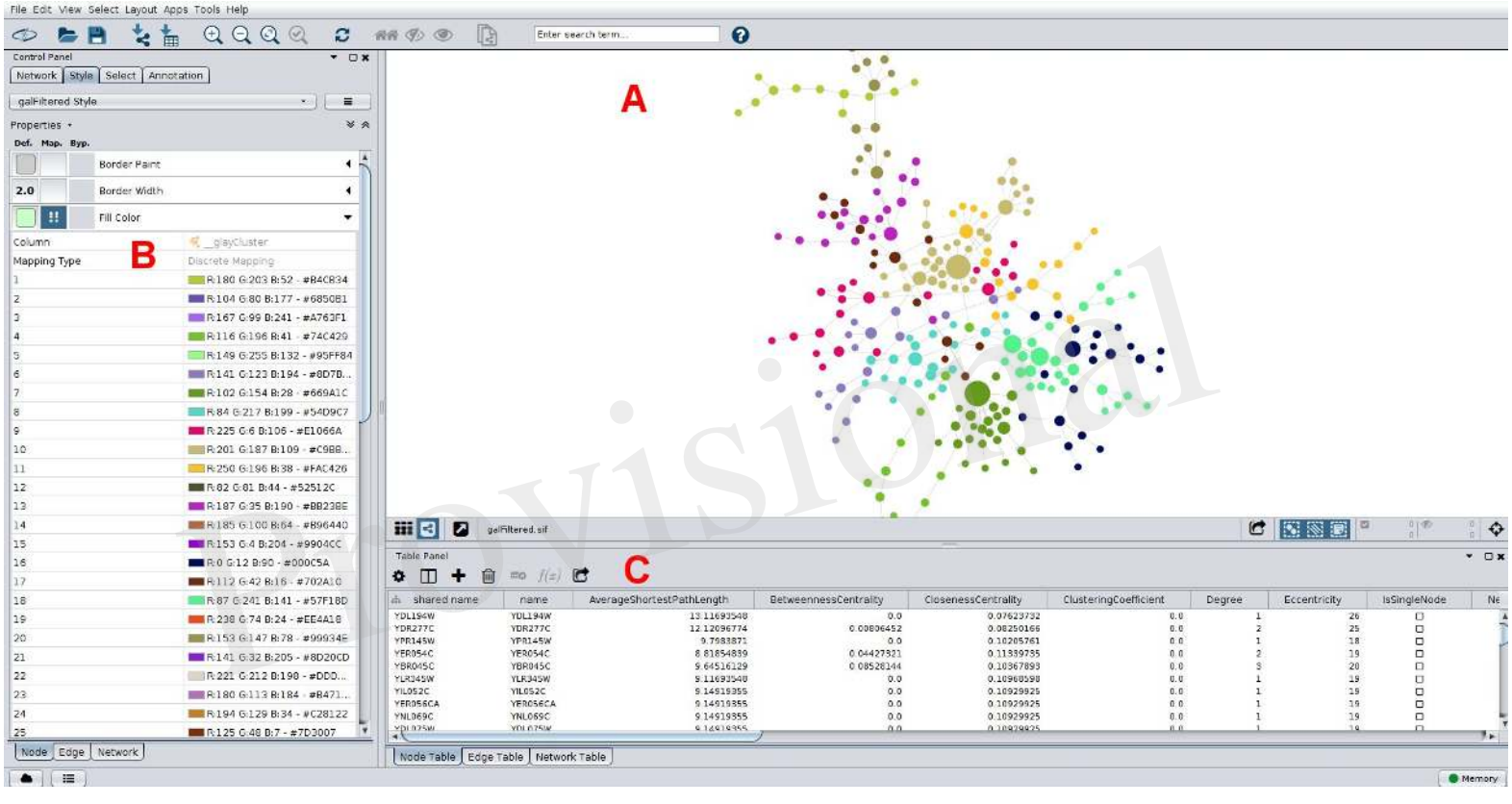


Figure 06.JPEG

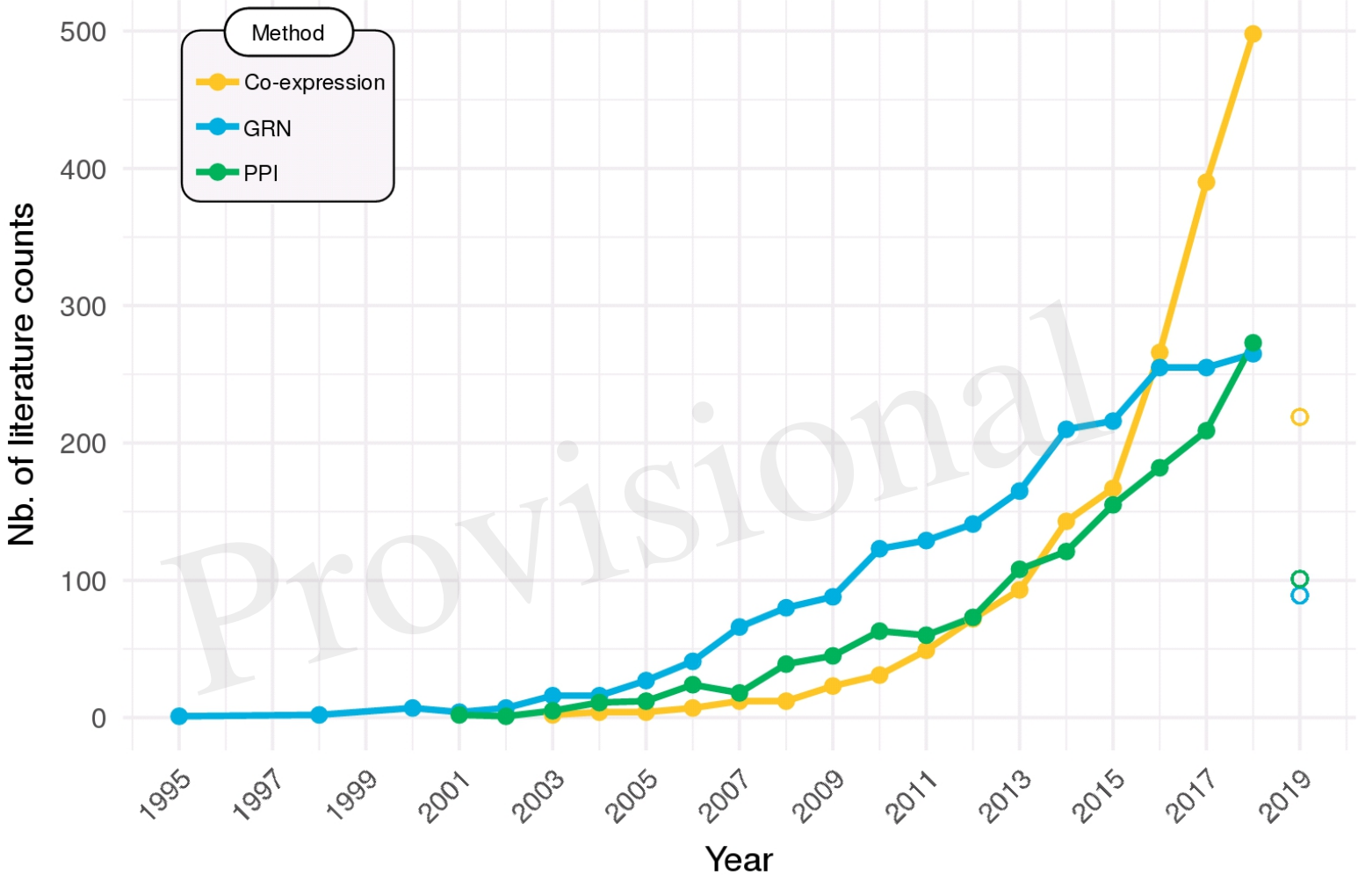


Figure 07.JPEG